PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N

A2

(11) International Publication Number: WO 98/07830

(43) International Publication Date: 26 February 1998 (26.02.98)

(21) International Application Number: PCT/US97/14900

(22) International Filing Date: 22 August 1997 (22.08.97)

(30) Priority Data:
60/024,428 22 August 1996 (22.08.96) US

(71) Applicants: THE INSTITUTE FOR GENOMIC RESEARCH [US/US]; 9712 Medical Center Drive, Rockville, MD 20850 (US). THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS [US/US]; 506 S. Wright Street, Urbana, IL 61802 (US). JOHNS HOPKINS UNIVERSITY SCHOOL OF MEDICINE [US/US]; Department of Molecular Biology and Genetics, Baltimore, MD 21205 (US).

(72) Inventors: BULT, Carol, J.; Box 525, Bar Harbor, ME 04609 (US). WHITE, Owen, R.; 886 Quince Orchard Boulevard # 202, Gaithersburg, MD 20878 (US). SMITH, Hamilton, O.; 8222 Carrbridge Circle, Baltimore, MD 21204 (US). WOESE, Carl, R.; 806 West Delaware Avenue, Urbana, IL 61801 (US). VENTER, J., Craig; 9708 Medical Center Drive, Rockville, MD 20850 (US).

(74) Agents: STEFFE, Eric, K. et al.; Sterne, Kessler, Goldstein & Fox P.L.L.C., Suite 600, 1100 New York Avenue, N.W., Washington, DC 20005-3934 (US).

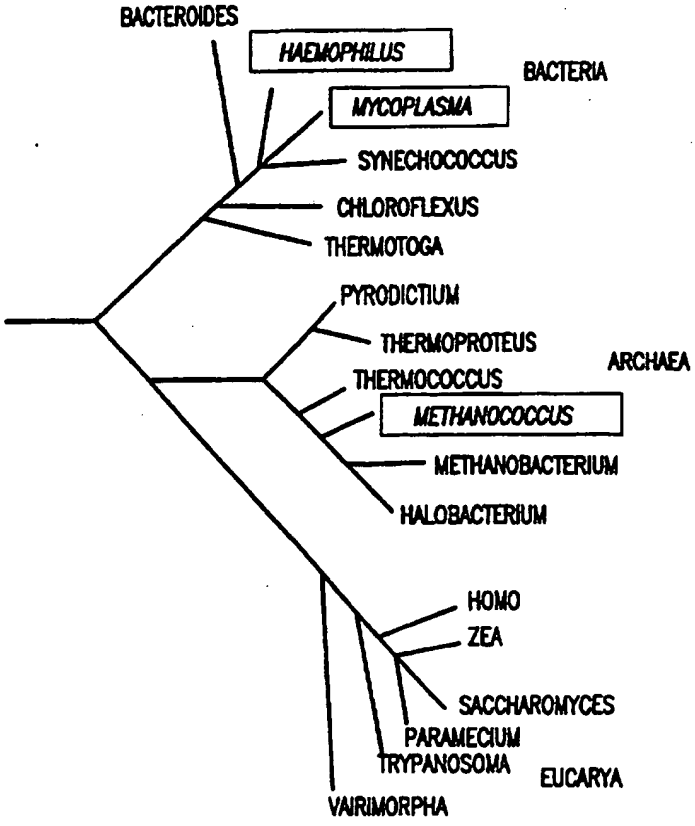
(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: COMPLETE GENOME SEQUENCE OF THE METHANOGENIC ARCHAEON, *METHANOCOCCUS JANNASCHII*

(57) Abstract

The present application describes the complete 1.66-megabase pair genome sequence of an autotrophic archaeon, *Methanococcus jannaschii*, and its 58- and 16-kilobase pair extrachromosomal elements. Also described are 1738 predicted protein-coding genes.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*

Background of the Invention

Statement as to Rights to Inventions Made Under Federally-Sponsored Research and Development

Part of the work performed during development of this invention utilized U.S. Government funds. The U.S. Government may have certain rights in the invention - DE-FC02-95ER61962; DE-FCO2-95ER61963; and NAGW 2554.

Field of the Invention

The present application discloses the complete 1.66-megabase pair genome sequence of an autotrophic archaeon, *Methanococcus jannaschii,* and its 58- and 16-kilobase pair extrachromosomal elements. Also identified are 1738 predicted protein-coding genes.

Related Background Art

The view of evolution in which all cellular organisms are in the first instance either prokaryotic or eukaryotic was challenged in 1977 by the finding that on the molecular level life comprises three primary groupings (Fox, G.E., *et al., Proc. Natl. Acad. Sci. USA 74*:4537 (1977); Woese, C.R. & Fox, G.E., *Proc. Natl. Acad. Sci. USA 74*:5088 (1977); Woese, C.R., *et al., Proc. Natl.,Acad. Sci. USA 87*:4576 (1990)): the eukaryotes (Eukarya) and two unrelated groups of prokaryotes, Bacteria and a new group now called the Archaea. Although Bacteria and Archaea are both prokaryotes in a cytological sense, they differ profoundly in their molecular makeup (Fox, G.E., *et al., Proc. Natl. Acad. Sci. USA 74*:4537 (1977); Woese, C.R. & Fox, G.E., *Proc. Natl. Acad. Sci. USA 74*:5088 (1977); Woese, C.R., *et al., Proc. Natl.,Acad. Sci. USA 87*:4576 (1990)).

Several lines of molecular evidence even suggest a specific relationship between Archaea and Eukarya (Iwabe, N., *et al.*, *Proc. Natl. Acad. Sci. USA 86*:9355 (1989); Gogarten J.P., *et al., Proc. Natl. Acad. Sci. USA 86*:6661 (1989); Brown, J.R. and Doolittle, W.F., *Proc. Natl. Acad. Sci. USA 92*:2441 (1995)).

The era of true comparative genomics has been ushered in by complete genome sequencing and analysis. We recently described the first two complete bacterial genome sequences, those of *Haemophilus influenzae* and *Mycoplasma genitalium* (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)). Large scale DNA sequencing efforts also have produced an extensive collection of sequence data from eukaryotes, including *Homo sapiens* (Adams, M.D., *et al.*, *Nature* 377:3 (1995)) and *Saccharomyces cerevisiae* (Levy, J., *Yeast 10*:1689 (1994)).

M. jannaschii was originally isolated by J.A. Leigh from a sediment sample collected from the sea floor surface at the base of a 2600 m deep "white smoker" chimney located at 21°N on the East Pacific Rise (Jones, W., *et al.*, *Arch. Microbiol. 136*:254 (1983)). *M. jannaschii* grows at pressures of up to more than 500 atm and over a temperature range of 48-94 °C, with an optimum temperature near 85 °C (Jones, W., *et al.*, *Arch. Microbiol. 136*:254 (1983)). The organism is autotrophic and a strict anaerobe; and, as the name implies, it produces methane. The dearth of archaeal nucleotide sequence data has hampered attempts to begin constructing a comprehensive comparative evolutionary framework for assessing the molecular basis of the origin and diversification of cellular life.

Summary of the Invention

The present invention is based on whole-genome random sequencing of an autotrophic archaeon, *Methanococcus jannaschii.* The *M. jannaschii* genome consists of three physically distinct elements: (i) a large circular chromosome; (ii) a large circular extrachromosomal element (ECE); and (iii) a small circular extrachromosomal element (ECE). The nucleotide sequences generated, the *M.*

jannaschii chromosome, the large ECE, and the small ECE, are respectively provided on pages 152-585 (SEQ ID NO:1), pages 585-600 (SEQ ID NO:2), and pages 601-605 (SEQ ID NO:3).

The present invention is further directed to isolated nucleic acid molecules comprising open reading frames (ORFs) encoding *M. jannaschii* proteins. The present invention also relates to variants of the nucleic acid molecules of the present invention, which encode portions, analogs or derivatives of *M. jannaschii* proteins. Further embodiments include isolated nucleic acid molecules comprising a polynucleotide having a nucleotide sequence at least 90% identical, and more preferably at least 95%, 96%, 97%, 98% or 99% identical, to the nucleotide sequence of a *M. jannaschii* ORF described herein.

The present invention also relates to recombinant vectors, which include the isolated nucleic acid molecules of the present invention, host cells containing the recombinant vectors, as well as methods for making such vectors and host cells for *M. jannaschii* protein production by recombinant techniques.

The invention further provides isolated polypeptides encoded by the *M. jannaschii* ORFs. It will be recognized that some amino acid sequences of the polypeptides described herein can be varied without significant effect on the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity. In general, it is possible to replace residues which form the tertiary structure, provided that residues performing a similar function are used. In other instances, the type of residue may be completely unimportant if the alteration occurs at a non-critical region of the protein.

In another aspect, the invention provides a peptide or polypeptide comprising an epitope-bearing portion of a polypeptide of the invention. The epitope-bearing portion is an immunogenic or antigenic epitope useful for raising antibodies.

Brief Description of the Figures

Figure 1. A schematic showing the relationship of the three domains of life based on sequence data from the small subunit of rRNA (Fox, G.E., *et al.*, *Proc. Natl. Acad. Sci. USA 74*:4537 (1977); Woese, C.R. & Fox, G.E., *Proc. Natl. Acad. Sci. USA 74*:5088 (1977); Woese, C.R., *et al.*, *Proc. Natl.,Acad. Sci. USA 87*:4576 (1990)).

Figure 2. Structure of a putative family of insertion sequence (IS) elements in the *M. jannaschii* genome. The family of elements has been named IS*AMJ1* and contains 11 members distributed among three groups (A, B, and C). The outer rectangle indicates the entire IS element; the interior rectangles indicate the predicted coding regions, oriented with the NH_2-termini to the left. DNA immediately adjacent to the NH_2-termini is 75 to 100% identical over 50 bp; DNA sequence similarity at the COOH-termini ends immediately after the stop codon. Black triangles indicate terminal inverted repeats. Fill patterns indicate which regions are missing from the elements in groups B and C. (A) Two copies of this family are 642 bp long and are 97% similar to each other at the nucleotide level. They appear to encode a protein 214 amino acids in length (ORFs MJ0017 and MJ1466) that are 27% identical to the IS*240* transposase of *Bacillus thuriengiensis* (GenBank Accession number: M23741). (B) Eight copies of the family range in length from 358 to 360 bp and are missing a 342-bp internal region relative to the two members of group A. Some members of group B have putative frameshifts (indicated by solid arrows) and in-frame UGA codons (indicated by open arrows). (C) The single copy in group C is 265 bp in length and occurs on the large ECE. The 436 bp internal region missing from this element is different than that of the members of group B.

Figure 3. Structure of a multicopy repetitive element in the *M. jannaschii* genome. Of the 18 copies identified on the main chromosome, seven are oriented in one direction (plus strand) and 11 are oriented in the opposite strand. Each element consists of a long, 391- to 425-bp repeat segment (designated LR) followed by up to 25 short, 27- to 28-bp repeat segments (designated SR). Each

SR segment is separated by 31 to 51 bp of sequence that is unique within and between each complete repeat element. (A) The longest repeat element has an LR segment followed by 25 SR segments, and spans more than 2 kbp, and (B) the shortest complete element has an LR segment followed by two SR segments. (C) One element is present in the genome with five SR segments and no LR component. (D and E) The LR segments of two elements in the genome are truncated at the end adjacent to the SR segments, both are followed by a single SR segment.

Figure 4. Block diagram of a computer system 102 that can be used to implement the computer-based systems of present invention.

Detailed Description of the Invention

The present invention is based on whole-genome random sequencing of an autotrophic archaeon, *Methanococcus jannaschii.* The *M. jannaschii* genome consists of three physically distinct elements: (i) a large circular chromosome of 1,664,976 base pairs (bp) (shown on pages 152-585 and in SEQ ID NO:1), which contains 1682 predicted protein-coding regions and has a G+C content of 31.4%; (ii) a large circular extrachromosomal element (the large ECE) of 58,407 bp (shown on pages 585-600 and in SEQ ID NO:2), which contains 44 predicted protein-coding regions and has a G+C content of 28.2%; and (iii) a small circular extrachromosomal element (the small ECE) of 16,550 bp (shown on pages 601-605 and in SEQ ID NO:3), which contains 12 predicted protein-coding regions and has a G+C content of 28.8%.

The primary nucleotide sequences generated, the *M. jannaschii* chromosome, the large ECE, and the small ECE, are provided in SEQ ID NOs:1, 2, and 3, respectively. As used herein, the "primary sequence" refers to the nucleotide sequence represented by the IUPAC nomenclature system. The present invention provides the nucleotide sequences of SEQ ID NOs:1, 2, and 3, or a representative fragment thereof, in a form which can be readily used, analyzed, and interpreted by a skilled artisan.

As used herein, a "representative fragment" refers to *M. jannaschii* protein-encoding regions (also referred to herein as open reading frames), expression modulating fragments, uptake modulating fragments, and fragments that can be used to diagnose the presence of *M. jannaschii* in a sample. A non-limiting identification of such representative fragments is provided in Tables 2(a) and 3. As described in detail below, representative fragments of the present invention further include nucleic acid molecules having a nucleotide sequence at least 90% identical, preferably at least 95, 96%, 97%, 98%, or 99% identical, to an ORF identified in Table 2(a) or 3.

As indicated above, the nucleotide sequence information provided in SEQ ID NOs:1, 2 and 3 was obtained by sequencing the *M. jannaschii* genome using a megabase shotgun sequencing method. The sequences provided in SEQ ID NOs:1, 2 and 3 are highly accurate, although not necessarily a 100% perfect, representation of the nucleotide sequence of the *M. jannaschii* genome. As discussed in detail below, using the information provided in SEQ ID NOs:1, 2 and 3 and in Tables 2(a) and 3 together with routine cloning and sequencing methods, one of ordinary skill in the art would be able to clone and sequence all "representative fragments" of interest including open reading frames (ORFs) encoding a large variety of *M. jannaschii* proteins. In rare instances, this may reveal a nucleotide sequence error present in the nucleotide sequences disclosed in SEQ ID NOs: 1, 2, and 3. Thus, once the present invention is made available (i.e., once the information in SEQ ID NOs:1, 2, and 3 and in Tables 2(a) and 3 have been made available), resolving a rare sequencing error would be well within the skill of the art. Nucleotide sequence editing software is publicly available. For example, Applied Biosystem's (AB) AutoAssembler™ can be used as an aid during visual inspection of nucleotide sequences.

Even if all of the rare sequencing errors were corrected, it is predicted that the resulting nucleotide sequences would still be at least about 99.9% identical to the reference nucleotide sequences in SEQ ID NOs:1, 2, and 3. Thus, the present invention further provides nucleotide sequences that are at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1, 2, or 3 in a form which can

be readily used, analyzed and interpreted by the skilled artisan. Methods for determining whether a nucleotide sequence is at least 99.9% identical to a reference nucleotide sequence of the present invention are described below.

Nucleic Acid Molecules

The present invention is directed to isolated nucleic acid fragments of the *M. jannaschii* genome. Such fragments include, but are not limited to, nucleic acid molecules encoding polypeptides (hereinafter open reading frames (ORFs)), nucleic acid molecules that modulate the expression of an operably linked ORF (hereinafter expression modulating fragments (EMFs)), nucleic acid molecules that mediate the uptake of a linked DNA fragment into a cell (hereinafter uptake modulating fragments (UMFs)), and nucleic acid molecules that can be used to diagnose the presence of *M. jannaschii* in a sample (hereinafter diagnostic fragments (DFs)).

By "isolated nucleic acid molecule(s)" is intended a nucleic acid molecule, DNA or RNA, that has been removed from its native environment. For example, recombinant DNA molecules contained in a vector are considered isolated for the purposes of the present invention. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells, purified (partially or substantially) DNA molecules in solution, and nucleic acid molecules produced synthetically. Isolated RNA molecules include *in vitro* RNA transcripts of the DNA molecules of the present invention.

In one embodiment, *M. jannaschii* DNA can be mechanically sheared to produce fragments about 15-20 kb in length, which can be used to generate a *M. jannaschii* DNA library by insertion into lambda clones as described in Example 1 below. Primers flanking an ORF described in Table 2(a) or 3 can then be generated using the nucleotide sequence information provided in SEQ ID NO:1, 2, or 3. The polymerase chain reaction (PCR) is then used to amplify and isolate the ORF from the lambda DNA library. PCR cloning is well known in the art. Thus, given SEQ ID NOs:1, 2, and 3, and Tables 2(a) and 3, it would be routine

to isolate any ORF or other representative fragment of the *M. jannaschii* genome. Isolated nucleic acid molecules of the present invention include, but are not limited to, single stranded and double stranded DNA, and single stranded RNA, and complements thereof.

Tables 2(a), 2(b) and 3 describe ORFs in the *M. jannaschii* genome. In particular, Table 2(a) (pages 67-115 below) indicates the location of ORFs (i.e., the position) within the *M. jannaschii* genome that putatively encode the recited protein based on homology matching with protein sequences from the organism appearing in parentheticals (see the fourth column of Table 2(a)). The first column of Table 2(a) provides a name for each ORF. The second and third columns in Table 2(a) indicate an ORF's position in the nucleotide sequence provided in SEQ ID NO:1, 2 or 3. One of ordinary skill in the art will appreciate that the ORFs may be oriented in opposite directions in the *M. jannaschii* genome. This is reflected in columns 2 and 3. The fifth column of Table 2(a) indicates the percent identity of the protein sequence encoded by an ORF to the corresponding protein sequence from the organism appearing in parentheticals in the fourth column. The sixth column of Table 2(a) indicates the percent similarity of the protein sequence encoded by an ORF to the corresponding protein sequence from the organism appearing in parentheticals in the fourth column. The concepts of percent identity and percent similarity of two polypeptide sequences are well understood in the art and are described in more detail below. The eighth column in Table 2(a) indicates the length of the ORF in nucleotides. Each identified gene has been assigned a putative cellular role category adapted from Riley (Riley, M., *Microbiol. Rev. 57*:862 (1993)).

Table 2(b) (page 116 below) provides the single ORF identified by the present inventors that matches a previously published *M. jannaschii* gene. In particular, ORF MJ0479, which is 585 nucleotides in length and is positioned at nucleotides 1,050,508 to 1,049,948 in SEQ ID NO:1, shares 100% identity to the previously published *M. jannaschii* adenylate kinase gene.

Table 3 (pages 117-150 below) provides ORFs of the *M. jannaschii* genome that did not elicit a homology match with a known sequence from either

M. jannaschii or another organism. As above, the first column in Table 3 provides the ORF name and the second and third columns indicate an ORF's position in SEQ ID NO:1, 2, or 3.

Table 4 (page 151 below) provides genes of *M. jannaschii* that contain inteins.

In the above-described Tables, there are three groups of ORF names. The one thousand six hundred and eighty two ORFs named "MJ-" (MJ0001-MJ1682) were identified on the *M. jannaschii* chromosome (SEQ ID NO:1). The forty four ORFs named "MJECL-" (MJECL01-MJECL44) were identified on the large ECE (SEQ ID NO:2). The twelve ORFs named "MJECS-" (MJECS01-MJES12) were identified on the small ECE (SEQ ID NO:3).

Further details concerning the algorithms and criteria used for homology searches are provided in the Examples below. A skilled artisan can readily identify ORFs in the *Methanococcus jannaschii* genome other than those listed in Tables 2(a), 2(b) and 3, such as ORFs that are overlapping or encoded by the opposite strand of an identified ORF in addition to those ascertainable using the computer-based systems of the present invention.

Isolated nucleic acid molecules of the present invention include DNA molecules having a nucleotide sequence substantially different than the nucleotide sequence of an ORF described in Table 2(a) or 3, but which, due to the degeneracy of the genetic code, still encode a *M. jannaschii* protein. The genetic code is well known in the art. Thus, it would be routine to generate such degenerate variants.

The present invention further relates to variants of the nucleic acid molecules of the present invention, which encode portions, analogs or derivatives of a *M. Jannaschii* protein encoded by an ORF described in Table 2(a) or 3. Non-naturally occurring variants may be produced using art-known mutagenesis techniques and include those produced by nucleotide substitutions, deletions or additions. The substitutions, deletions or additions may involve one or more nucleotides. The variants may be altered in coding regions, non-coding regions, or both. Alterations in the coding regions may produce conservative or

non-conservative amino acid substitutions, deletions or additions. Especially preferred among these are silent substitutions, additions and deletions, which do not alter the properties and activities of the *M. jannaschii* protein or portions thereof. Also especially preferred in this regard are conservative substitutions.

Further embodiments of the invention include isolated nucleic acid molecules comprising a polynucleotide having a nucleotide sequence at least 90% identical, and more preferably at least 95%, 96%, 97%, 98% or 99% identical, to (a) the nucleotide sequence of an ORF described in Table 2(a) or 3, (b) the nucleotide sequence of an ORF described in Table 2(a) or 3, but lacking the codon for the N-terminal methionine residue, if present, or (c) a nucleotide sequence complementary to any of the nucleotide sequences in (a) or (b). By a polynucleotide having a nucleotide sequence at least, for example, 95% identical to the reference *M. jannaschii* ORF nucleotide sequence is intended that the nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the ORF sequence. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference ORF nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. These mutations of the reference sequence may occur at the 5′ or 3′ terminal positions of the reference nucleotide sequence or anywhere between those terminal positions, interspersed either individually among nucleotides in the reference sequence or in one or more contiguous groups within the reference sequence.

As a practical matter, whether any particular nucleic acid molecule is at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleotide sequence of a *M. jannaschii* ORF can be determined conventionally using known computer programs such as the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). Bestfit uses the local homology algorithm

of Smith and Waterman, *Advances in Applied Mathematics 2*: 482-489 (1981), to find the best segment of homology between two sequences. When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference nucleotide sequence and that gaps in homology of up to 5% of the total number of nucleotides in the reference sequence are allowed.

Preferred are nucleic acid molecules having sequences at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleic acid sequence of a *M. jannaschii* ORF that encode a functional polypeptide. By a "functional polypeptide" is intended a polypeptide exhibiting activity similar, but not necessarily identical, to an activity of the protein encoded by the *M. jannaschii* ORF. For example, the *M. jannaschii* ORF MJ1434 encodes an endonuclease that degrades DNA. Thus, a "functional polypeptide" encoded by a nucleic acid molecule having a nucleotide sequence, for example, 95% identical to the nucleotide sequence of MJ1434, will also degrade DNA. As the skilled artisan will appreciate, assays for determining whether a particular polypeptide is "functional" will depend on which ORF is used as the reference sequence. Depending on the reference ORF, the assay chosen for measuring polypeptide activity will be readily apparent in light of the role categories provided in Table 2(a).

Of course, due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large number of the nucleic acid molecules having a sequence at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequence of a reference ORF will encode a functional polypeptide. In fact, since degenerate variants all encode the same amino acid sequence, this will be clear to the skilled artisan even without performing a comparison assay for protein activity. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate variants, a reasonable number will also encode a functional polypeptide. This is because the skilled artisan is fully aware of amino acid substitutions that are either less likely or not

likely to significantly affect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid).

For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie, J. U. *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990), wherein the authors indicate that there are two main approaches for studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. The second approach uses genetic engineering to introduce amino acid changes at specific positions of a cloned gene and selections or screens to identify sequences that maintain functionality. As the authors state, these studies have revealed that proteins are surprisingly tolerant of amino acid substitutions. The authors further indicate which amino acid changes are likely to be permissive at a certain position of the protein. For example, most buried amino acid residues require nonpolar side chains, whereas few features of surface side chains are generally conserved. Other such phenotypically silent substitutions are described in Bowie, J.U. *et al., supra,* and the references cited therein.

The present invention is further directed to fragments of the isolated nucleic acid molecules described herein. By a fragment of an isolated nucleic acid molecule having the nucleotide sequence of a *M. jannaschii* ORF is intended fragments at least about 15 nt, and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably, at least about 40 nt in length that are useful as diagnostic probes and primers as discussed herein. Of course, larger fragments 50-500 nt in length are also useful according to the present invention as are fragments corresponding to most, if not all, of the nucleotide sequence of a *M. jannaschii* ORF. By a fragment at least 20 nt in length, for example, is intended fragments that include 20 or more contiguous bases from the nucleotide sequence of a *M. jannaschii* ORF. Since *M. jannaschii* ORFs are listed in Tables 2(a) and 3 and the genome sequence has been provided, generating such DNA fragments would be routine to the skilled artisan. For

example, restriction endonuclease cleavage or shearing by sonication could easily be used to generate fragments of various sizes. Alternatively, such fragments could be generated synthetically.

Preferred nucleic acid fragments of the present invention include nucleic acid molecules encoding epitope-bearing portions of a *M. jannaschii* protein. Methods for determining such epitope-bearing portions are described in detail below.

In another aspect, the invention provides an isolated nucleic acid molecule comprising a polynucleotide that hybridizes under stringent hybridization conditions to a portion of the polynucleotide in a nucleic acid molecule of the invention described above, for instance, an ORF described in Table 2(a) or 3. By "stringent hybridization conditions" is intended overnight incubation at 42°C in a solution comprising: 50% formamide, 5x SSC (150 mM NaCl, 15mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5x Denhardt's solution, 10% dextran sulfate, and 20 g/ml denatured, sheared salmon sperm DNA, followed by washing the filters in 0.1x SSC at about 65°C.

By a polynucleotide that hybridizes to a "portion" of a polynucleotide is intended a polynucleotide (either DNA or RNA) hybridizing to at least about 15 nucleotides (nt), and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably about 30-70 nt of the reference polynucleotide. These are useful as diagnostic probes and primers as discussed above and in more detail below.

Of course, polynucleotides hybridizing to a larger portion of the reference polynucleotide (e.g., a *M. jannaschii* ORF), for instance, a portion 50-500 nt in length, or even to the entire length of the reference polynucleotide, are also useful as probes according to the present invention, as are polynucleotides corresponding to most, if not all, of a *M. jannaschii* ORF.

By "expression modulating fragment" (EMF), is intended a series of nucleotides that modulate the expression of an operably linked ORF or EMF. A sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs are fragments that induce the expression of an operably linked ORF in response to a specific regulatory factor or physiological event. EMF sequences can be identified within the *M. jannaschii* genome by their proximity to the ORFs described in Tables 2(a), 2(b), and 3. An intergenic segment, or a fragment of the intergenic segment, from about 10 to 200 nucleotides in length, taken 5′ from any one of the ORFs of Tables 2(a), 2(b) or 3 will modulate the expression of an operably linked 3′ ORF in a fashion similar to that found with the naturally linked ORF sequence. As used herein, an "intergenic segment" refers to the fragments of the *M. jannaschii* genome that are between two ORF(s) herein described. Alternatively, EMFs can be identified using known EMFs as a target sequence or target motif in the computer-based systems of the present invention.

The presence and activity of an EMF can be confirmed using an EMF trap vector. An EMF trap vector contains a cloning site 5′ to a marker sequence. A marker sequence encodes an identifiable phenotype, such as antibiotic resistance or a complementing nutrition auxotrophic factor, which can be identified or assayed when the EMF trap vector is placed within an appropriate host under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence. A more detailed discussion of various marker sequences is provided below.

A sequence that is suspected as being an EMF is cloned in all three reading frames in one or more restriction sites upstream from the marker sequence in the EMF trap vector. The vector is then transformed into an appropriate host using known procedures and the phenotype of the transformed host in examined under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence.

By "uptake modulating fragment" (UMF), is intended a series of nucleotides that mediate the uptake of a linked DNA fragment into a cell. UMFs can be readily identified using known UMFs as a target sequence or target motif with the computer-based systems described below. The presence and activity of a UMF can be confirmed by attaching the suspected UMF to a marker sequence. The resulting nucleic acid molecule is then incubated with an appropriate host under appropriate conditions and the uptake of the marker sequence is determined. As described above, a UMF will increase the frequency of uptake of a linked marker sequence.

By a "diagnostic fragment" (DF), is intended a series of nucleotides that selectively hybridize to *M. jannaschii* sequences. DFs can be readily identified by identifying unique sequences within the *M. jannaschii* genome, or by generating and testing probes or amplification primers consisting of the DF sequence in an appropriate diagnostic format for amplification or hybridization selectivity.

Each of the ORFs of the *M. jannaschii* genome disclosed in Tables 2(a) and 3, and the EMF found 5′ to the ORF, can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes or diagnostic amplification primers to detect the presence *M. jannaschii* in a sample. This is especially the case with the fragments or ORFs of Table 3, which will be highly selective for *M. jannaschii*.

In addition, the fragments of the present invention, as broadly described, can be used to control gene expression through triple helix formation or antisense DNA or RNA, both of which methods are based on the binding of a polynucleotide sequence to DNA or RNA. Polynucleotides suitable for use in these methods are usually 20 to 40 bases in length and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*:1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)).

Triple helix- formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide.

Vectors and Host Cells

The present invention further provides recombinant constructs comprising one or more fragments of the *M. jannaschii* genome. The recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which, for example, a *M. jannaschii* ORF is inserted. The vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the ORF. For vectors comprising the EMFs and UMFs of the present invention, the vector may further comprise a marker sequence or heterologous ORF operably linked to the EMF or UMF. Large numbers of suitable vectors and promoters are known to those of skill in the art and are commercially available for generating the recombinant constructs of the present invention. The following vectors are provided by way of example. Bacterial: pBs, phagescript, PsiX174, pBluescript SK, pBs KS, pNH8a, pNH16a, pNH18a, pNH46a (Stratagene); pTrc99A, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLneo, pSV2cat, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia).

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, and trc. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the

appropriate vector and promoter is well within the level of ordinary skill in the art.

The present invention further provides host cells containing any one of the isolated fragments (preferably an ORF) of the *M. jannaschii* genome described herein. The host cell can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or the host cell can be a procaryotic cell, such as a bacterial cell. Introduction of the recombinant construct into the host cell can be effected by calcium phosphate transfection, DEAE, dextran mediated transfection, or electroporation (Davis, L. *et al.*, *Basic Methods in Molecular Biology* (1986)). Host cells containing, for example, a *M. jannaschii* ORF can be used conventionally to produce the encoded protein.

Polypeptides and Fragments

The invention further provides an isolated polypeptide encoded by a *M. jannaschii* ORF described in Tables 2(a) or 3, or a peptide or polypeptide comprising a portion of the isolated polypeptide. The terms "peptide" and "oligopeptide" are considered synonymous (as is commonly recognized) and each term can be used interchangeably as the context requires to indicate a chain of at least two amino acids coupled by peptidyl linkages. The word "polypeptide" is used herein for chains containing more than ten amino acid residues.

It will be recognized in the art that some amino acid sequence of the *M. jannaschii* polypeptide can be varied without significant affect of the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity. In general, it is possible to replace residues which form the tertiary structure, provided that residues performing a similar function are used. In other instances, the type of residue may be completely unimportant if the alteration occurs at a non-critical region of the protein.

Thus, the invention further includes variations of a *M. jannaschii* protein encoded by an ORF described in Table 2(a) or 3 that show substantial protein

activity. Methods for assaying such "functional polypeptides" for protein activity are described above. Variations include deletions, insertions, inversions, repeats, and type substitutions (for example, substituting one hydrophilic residue for another, but not strongly hydrophilic for strongly hydrophobic as a rule). Small changes or such "neutral" amino acid substitutions will generally have little effect on protein activity.

Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu and Ile; interchange of the hydroxyl residues Ser and Thr, exchange of the acidic residues Asp and Glu, substitution between the amide residues Asn and Gln, exchange of the basic residues Lys and Arg and replacements among the aromatic residues Phe, Tyr.

As indicated in detail above, further guidance concerning amino acid changes that are likely to be phenotypically silent (i.e., are not likely to have a significant deleterious effect on function) can be found in Bowie, J.U., *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990).

The fragment, derivative, variant or analog of a *M. jannaschii* polypeptide encoded by an ORF described in Table 2(a) or 3, may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the polypeptide, such as an IgG Fc fusion region peptide or leader or secretory sequence or a sequence which is employed for purification of the polypeptide or a proprotein sequence. Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

Of particular interest are substitutions of charged amino acids with another charged amino acid and with neutral or negatively charged amino acids. The latter results in proteins with reduced positive charge to improve the characteristics of a *M. jannaschii* ORF-encoded protein. The prevention of aggregation is highly desirable. Aggregation of proteins not only results in a loss of activity but can also be problematic when preparing pharmaceutical formulations, because they can be immunogenic. (Pinckard *et al., Clin. Exp. Immunol. 2*:331-340 (1967); Robbins *et al., Diabetes 36*:838-845 (1987); Cleland *et al. Crit. Rev. Therapeutic Drug Carrier Systems 10*:307-377 (1993)).

As indicated, changes are preferably of a minor nature, such as conservative amino acid substitutions that do not significantly affect the folding or activity of the protein (see Table 1).

TABLE 1. Conservative Amino Acid Substitutions.

Aromatic	Phenylalanine Tryptophan Tyrosine
Hydrophobic	Leucine Isoleucine Valine
Polar	Glutamine Asparagine
Basic	Arginine Lysine Histidine
Acidic	Aspartic Acid Glutamic Acid
Small	Alanine Serine Threonine Methionine Glycine

Amino acids in a *M. jannaschii* ORF-encoded protein of the present invention that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis

(Cunningham and Wells, *Science 244*:1081-1085 (1989)). The latter procedure introduces single alanine mutations at every residue in the molecule.

The polypeptides of the present invention are preferably provided in an isolated form. By "isolated polypeptide" is intended a polypeptide removed from its native environment. Thus, a polypeptide produced and/or contained within a recombinant host cell is considered isolated for purposes of the present invention. Also intended as an "isolated polypeptide" are polypeptides that have been purified, partially or substantially, from a recombinant host cell. For example, a recombinantly produced version of a *M. jannaschii* ORF-encoded protein can be substantially purified by the one-step method described in Smith and Johnson, *Gene 67*:31-40 (1988).

The polypeptides of the present invention include the proteins encoded by (a) an ORF described in Table 2(a) or 3 or (b) an ORF described in Table 2(a) or 3, but minus the codon for the N-terminal methionine residue, if present, as well as polypeptides that have at least 90% similarity, more preferably at least 95% similarity, and still more preferably at least 96%, 97%, 98% or 99% similarity to a *M. jannaschii* ORF-encoded protein. Further polypeptides of the present invention include polypeptides at least 90% identical, more preferably at least 95% identical, still more preferably at least 96%, 97%, 98% or 99% identical to a *M. jannaschii* ORF-encoded protein.

By "% similarity" for two polypeptides is intended a similarity score produced by comparing the amino acid sequences of the two polypeptides using the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711) and the default settings for determining similarity. Bestfit uses the local homology algorithm of Smith and Waterman *(Advances in Applied Mathematics 2*:482-489, 1981) to find the best segment of similarity between two sequences.

By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a reference amino acid sequence of a *M. jannaschii* ORF-encoded protein is intended that the amino acid sequence of the polypeptide is

identical to the reference sequence except that the polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the reference sequence. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a reference amino acid sequence, up to 5% of the amino acid residues in the reference sequence may be deleted or substituted with another amino acid, or a number of amino acids up to 5% of the total amino acid residues in the reference sequence may be inserted into the reference sequence. These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence.

As a practical matter, whether any particular polypeptide has an amino acid sequence at least 90%, 95%, 96%, 97%, 98% or 99% identical to the amino acid sequence of a *M. jannaschii* ORF-encoded protein can be determined conventionally using known computer programs such the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference amino acid sequence and that gaps in homology of up to 5% of the total number of amino acid residues in the reference sequence are allowed.

As described in detail below, the polypeptides of the present invention can also be used to raise polyclonal and monoclonal antibodies, which are useful in assays for detecting *M. jannaschii* protein expression.

In another aspect, the invention provides a peptide or polypeptide comprising an epitope-bearing portion of a polypeptide of the invention. The epitope of this polypeptide portion is an immunogenic or antigenic epitope of a polypeptide of the invention. An "immunogenic epitope" is defined as a part of

a protein that elicits an antibody response when the whole protein is the immunogen. These immunogenic epitopes are believed to be confined to a few loci on the molecule. On the other hand, a region of a protein molecule to which an antibody can bind is defined as an "antigenic epitope." The number of immunogenic epitopes of a protein generally is less than the number of antigenic epitopes. See, for instance, Geysen *et al., Proc. Natl. Acad. Sci. USA 81*:3998-4002 (1983).

As to the selection of peptides or polypeptides bearing an antigenic epitope (i.e., that contain a region of a protein molecule to which an antibody can bind), it is well known in that art that relatively short synthetic peptides that mimic part of a protein sequence are routinely capable of eliciting an antiserum that reacts with the partially mimicked protein. See, for instance, Sutcliffe, J. G., Shinnick, T. M., Green, N. and Learner, R.A. (1983). Antibodies that react with predetermined sites on proteins are described in *Science 219:*660-666. Peptides capable of eliciting protein-reactive sera are frequently represented in the primary sequence of a protein, can be characterized by a set of simple chemical rules, and are confined neither to immunodominant regions of intact proteins (i.e., immunogenic epitopes) nor to the amino or carboxyl terminals. Peptides that are extremely hydrophobic and those of six or fewer residues generally are ineffective at inducing antibodies that bind to the mimicked protein; longer, peptides, especially those containing proline residues, usually are effective. Sutcliffe *et al., supra*, at 661. For instance, 18 of 20 peptides designed according to these guidelines, containing 8-39 residues covering 75% of the sequence of the influenza virus hemagglutinin HA1 polypeptide chain, induced antibodies that reacted with the HA1 protein or intact virus; and 12/12 peptides from the MuLV polymerase and 18/18 from the rabies glycoprotein induced antibodies that precipitated the respective proteins.

Antigenic epitope-bearing peptides and polypeptides of the invention are therefore useful to raise antibodies, including monoclonal antibodies, that bind specifically to a polypeptide of the invention. Thus, a high proportion of hybridomas obtained by fusion of spleen cells from donors immunized with an

antigen epitope-bearing peptide generally secrete antibody reactive with the native protein. Sutcliffe *et al.*, *supra*, at 663. The antibodies raised by antigenic epitope-bearing peptides or polypeptides are useful to detect the mimicked protein, and antibodies to different peptides may be used for tracking the fate of various regions of a protein precursor which undergoes post-translational processing. The peptides and anti-peptide antibodies may be used in a variety of qualitative or quantitative assays for the mimicked protein, for instance in competition assays since it has been shown that even short peptides (e.g., about 9 amino acids) can bind and displace the larger peptides in immunoprecipitation assays. See, for instance, Wilson *et al.*, *Cell 37*:767-778 (1984) at 777. The anti-peptide antibodies of the invention also are useful for purification of the mimicked protein, for instance, by adsorption chromatography using methods well known in the art.

Antigenic epitope-bearing peptides and polypeptides of the invention designed according to the above guidelines preferably contain a sequence of at least seven, more preferably at least nine and most preferably between about 15 to about 30 amino acids contained within the amino acid sequence of a polypeptide of the invention. However, peptides or polypeptides comprising a larger portion of an amino acid sequence of a polypeptide of the invention, containing about 30 to about 50 amino acids, or any length up to and including the entire amino acid sequence of a polypeptide of the invention, also are considered epitope-bearing peptides or polypeptides of the invention and also are useful for inducing antibodies that react with the mimicked protein. Preferably, the amino acid sequence of the epitope-bearing peptide is selected to provide substantial solubility in aqueous solvents (i.e., the sequence includes relatively hydrophilic residues and highly hydrophobic sequences are preferably avoided); and sequences containing proline residues are particularly preferred.

The epitope-bearing peptides and polypeptides of the invention may be produced by any conventional means for making peptides or polypeptides including recombinant means using nucleic acid molecules of the invention. For instance, a short epitope-bearing amino acid sequence may be fused to a larger

polypeptide which acts as a carrier during recombinant production and purification, as well as during immunization to produce anti-peptide antibodies. Epitope-bearing peptides also may be synthesized using known methods of chemical synthesis. For instance, Houghten has described a simple method for synthesis of large numbers of peptides, such as 10-20 mg of 248 different 13 residue peptides representing single amino acid variants of a segment of the HA1 polypeptide which were prepared and characterized (by ELISA-type binding studies) in less than four weeks. Houghten, R. A. (1985) General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proc. Natl. Acad. Sci. USA 82*:5131-5135. This "Simultaneous Multiple Peptide Synthesis (SMPS)" process is further described in U.S. Patent No. 4,631,211 to Houghten *et al.* (1986). In this procedure the individual resins for the solid-phase synthesis of various peptides are contained in separate solvent-permeable packets, enabling the optimal use of the many identical repetitive steps involved in solid-phase methods. A completely manual procedure allows 500-1000 or more syntheses to be conducted simultaneously. Houghten *et al.*, *supra*, at 5134.

Epitope-bearing peptides and polypeptides of the invention are used to induce antibodies according to methods well known in the art. See, for instance, Sutcliffe *et al.*, *supra*; Wilson *et al.*, *supra*; Chow, M. *et al.*, *Proc. Natl. Acad. Sci. USA 82*:910-914; and Bittle, F. J. *et al.*, *J. Gen. Virol. 66*:2347-2354 (1985). Generally, animals may be immunized with free peptide; however, anti-peptide antibody titer may be boosted by coupling of the peptide to a macromolecular carrier, such as keyhole limpet hemacyanin (KLH) or tetanus toxoid. For instance, peptides containing cysteine may be coupled to carrier using a linker such as m-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS), while other peptides may be coupled to carrier using a more general linking agent such as glutaraldehyde. Animals such as rabbits, rats and mice are immunized with either free or carrier-coupled peptides, for instance, by intraperitoneal and/or intradermal injection of emulsions containing about 100 g peptide or carrier protein and Freund's adjuvant. Several booster injections may be needed, for

instance, at intervals of about two weeks, to provide a useful titer of anti-peptide antibody which can be detected, for example, by ELISA assay using free peptide adsorbed to a solid surface. The titer of anti-peptide antibodies in serum from an immunized animal may be increased by selection of anti-peptide antibodies, for instance, by adsorption to the peptide on a solid support and elution of the selected antibodies according to methods well known in the art.

Immunogenic epitope-bearing peptides of the invention, i.e., those parts of a protein that elicit an antibody response when the whole protein is the immunogen, are identified according to methods known in the art. For instance, Geysen *et al.*, *supra*, discloses a procedure for rapid concurrent synthesis on solid supports of hundreds of peptides of sufficient purity to react in an enzyme-linked immunosorbent assay. Interaction of synthesized peptides with antibodies is then easily detected without removing them from the support. In this manner a peptide bearing an immunogenic epitope of a desired protein may be identified routinely by one of ordinary skill in the art. For instance, the immunologically important epitope in the coat protein of foot-and-mouth disease virus was located by Geysen *et al.* with a resolution of seven amino acids by synthesis of an overlapping set of all 208 possible hexapeptides covering the entire 213 amino acid sequence of the protein. Then, a complete replacement set of peptides in which all 20 amino acids were substituted in turn at every position within the epitope were synthesized, and the particular amino acids conferring specificity for the reaction with antibody were determined. Thus, peptide analogs of the epitope-bearing peptides of the invention can be made routinely by this method. U.S. Patent No. 4,708,781 to Geysen (1987) further describes this method of identifying a peptide bearing an immunogenic epitope of a desired protein.

Further still, U.S. Patent No. 5,194,392 to Geysen (1990) describes a general method of detecting or determining the sequence of monomers (amino acids or other compounds) which is a topological equivalent of the epitope (i.e., a "mimotope") which is complementary to a particular paratope (antigen binding site) of an antibody of interest. More generally, U.S. Patent No. 4,433,092 to Geysen (1989) describes a method of detecting or determining a sequence of

monomers which is a topographical equivalent of a ligand which is complementary to the ligand binding site of a particular receptor of interest. Similarly, U.S. Patent No. 5,480,971 to Houghten, R. A. *et al.* (1996) on Peralkylated Oligopeptide Mixtures discloses linear C_1-C_7-alkyl peralkylated oligopeptides and sets and libraries of such peptides, as well as methods for using such oligopeptide sets and libraries for determining the sequence of a peralkylated oligopeptide that preferentially binds to an acceptor molecule of interest. Thus, non-peptide analogs of the epitope-bearing peptides of the invention also can be made routinely by these methods.

The entire disclosure of each document cited in this section on "Polypeptides and Peptides" is hereby incorporated herein by reference.

As one of skill in the art will appreciate, the polypeptides of the present invention and the epitope-bearing fragments thereof described above can be combined with parts of the constant domain of immunoglobulins (IgG), resulting in chimeric polypeptides. These fusion proteins facilitate purification and show an increased half-life *in vivo*. This has been demonstrated, e.g., for chimeric proteins consisting of the first two domains of the human CD4-polypeptide and various domains of the constant regions of the heavy or light chains of mammalian immunoglobulins (EPA 394,827; Traunecker *et al.*, *Nature 331*:84-86 (1988)). Fusion proteins that have a disulfide-linked dimeric structure due to the IgG part can also be more efficient in binding and neutralizing other molecules than the monomeric protein or protein fragment alone (Fountoulakis *et al.*, *J Biochem 270*:3958-3964 (1995)).

Protein Function

Each ORF described in Table 2(a) was assigned to biological role categories adapted from Riley, M., *Microbiology Reviews 57(4)*:862 (1993)). This allows the skilled artisan to determine a function for each identified coding sequence. For example, a partial list of the *M. jannaschii* protein functions provided in Table 2(a) includes: methanogenesis, amino acid biosynthesis, cell

division, detoxification, protein secretion, transformation, central intermediary metabolism, energy metabolism, degradation of DNA, DNA replication, restriction, modification, recombination and repair, transcription, RNA processing, translation, degradation of proteins, peptides and glycopeptides, ribosomal proteins, translation factors, transport, tRNA modification, and drug and analog sensitivity. A more detailed description of several of these functions is provided in Example 1 below.

Diagnostic Assays

The present invention further provides methods to identify the expression of an ORF of the present invention, or homolog thereof, in a test sample, using one of the DFs or antibodies of the present invention. Such methods involve incubating a test sample with one or more of the antibodies or one or more of the DFs of the present invention and assaying for binding of the DFs or antibodies to components within the test sample.

Conditions for incubating a DF or antibody with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the DF or antibody used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification or immunological assay formats can readily be adapted to employ the DFs or antibodies of the present invention. Examples of such assays can be found in Chard, T., *An Introduction to Radioimmunoassay and Related Techniques*, Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock, G.R. *et al.*, *Techniques in Immunocytochemistry*, Academic Press, Orlando, FL Vol. 1 (1982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, P., *Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1985).

The test samples of the present invention include cells, protein or membrane extracts of cells. The test sample used in the above-described method

will vary based on the assay format, nature of the detection method and the cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are well known in the art and can be readily be adapted in order to obtain a sample which is compatible with the system utilized.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the assays of the present invention. Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers including comprising: (a) a first container comprising one of the DFs or antibodies of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of a bound DF or antibody.

A compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers or strips of plastic or paper. Such containers allow one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the antibodies used in the assay, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and containers which contain the reagents used to detect the bound antibody or DF.

Types of detection reagents include labeled nucleic acid probes, labeled secondary antibodies, or in the alternative, if the primary antibody is labeled, the enzymatic, or antibody binding reagents that are capable of reacting with the labeled antibody. One skilled in the art will readily recognize that the disclosed DFs and antibodies of the present invention can be readily incorporated into one of the established kit formats that are well known in the art.

Screening Assay for Binding Agents

Using the isolated proteins described herein, the present invention further provides methods of obtaining and identifying agents that bind to a protein encoded by a *M. jannaschii* ORF or to a fragment thereof.

The method involves:

(a) contacting an agent with an isolated protein encoded by a *M. jannaschii* ORF, or an isolated fragment thereof; and

(b) determining whether the agent binds to said protein or said fragment.

The agents screened in the above assay can be, but are not limited to, peptides, carbohydrates, vitamin derivatives, or other pharmaceutical agents. The agents can be selected and screened at random or rationally selected or designed using protein modeling techniques. For random screening, agents such as peptides, carbohydrates, pharmaceutical agents and the like are selected at random and are assayed for their ability to bind to the protein encoded by an ORF of the present invention.

Alternatively, agents may be rationally selected or designed. As used herein, an agent is said to be "rationally selected or designed" when the agent is chosen based on the configuration of the particular protein. For example, one skilled in the art can readily adapt currently available procedures to generate peptides, pharmaceutical agents and the like capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide peptides, for example see Hurby *et al.*, Application of Synthetic Peptides: Antisense Peptides, In *Synthetic Peptides, A User's Guide*, W.H. Freeman, NY (1992), pp. 289-307, and Kaspczak *et al., Biochemistry 28*:9230-8 (1989), or pharmaceutical agents, or the like.

In addition to the foregoing, one class of agents of the present invention, can be used to control gene expression through binding to one of the ORFs or EMFs of the present invention. As described above, such agents can be randomly

screened or rationally designed and selected. Targeting the ORF or EMF allows a skilled artisan to design sequence specific or element specific agents, modulating the expression of either a single ORF or multiple ORFs that rely on the same EMF for expression control.

One class of DNA binding agents are those that contain nucleotide base residues that hybridize or form a triple helix by binding to DNA or RNA. Such agents can be based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives having base attachment capacity.

Agents suitable for use in these methods usually contain 20 to 40 bases and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*: 1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)). Triple helix-formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide and other DNA binding agents.

Computer Related Embodiments

The nucleotide sequence provided in SEQ ID NO:1, 2, or 3, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to the sequence provided in SEQ ID NO:1, 2, or 3, can be "provided" in a variety of mediums to facilitate use thereof. As used herein, provided refers to a manufacture, other than an isolated nucleic acid molecule, that contains a nucleotide sequence of the present invention, i.e., the nucleotide sequence provided in SEQ ID NO:1, 2, or 3, a representative fragment thereof, or a

nucleotide sequence at least 99.9% identical to SEQ ID NO:1, 2, or 3. Such a manufacture provides the *M. jannaschii* genome or a subset thereof (e.g., a *M. jannaschii* open reading frame (ORF)) in a form that allows a skilled artisan to examine the manufacture using means not directly applicable to examining the *M. jannaschii* genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of dataprocessor structuring formats

(e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing the nucleotide sequence of SEQ ID NO:1, 2, or 3, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1, 2, or 3, in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al.*, *Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the *M. jannaschii* genome that contain homology to ORFs or proteins from other organisms. Such ORFs are protein-encoding fragments within the *M. jannaschii* genome and are useful in producing commercially important proteins such as enzymes used in methanogenesis, amino acid biosynthesis, metabolism, fermentation, transcription, translation, RNA processing, nucleic acid and protein degradation, protein modification, and DNA replication, restriction, modification, recombination, and repair. A comprehensive list of ORFs encoding commercially important *M. jannaschii* proteins is provided in Tables 2(a) and 3.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the *M. jannaschii* genome. As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As indicated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory that can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention. As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the *M. jannaschii* genome that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available and can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that during searches for commercially important fragments of the *M. jannaschii* genome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is

formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

Thus, the present invention further provides an input means for receiving a target sequence, a data storage means for storing the target sequence and the homologous *M. jannaschii* sequence identified using a search means as described above, and an output means for outputting the identified homologous *M. jannaschii* sequence. A variety of structural formats for the input and output means can be used to input and output information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *M. jannaschii* genome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *M. jannaschii* genome. For example, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) can be used to identify open reading frames within the *M. jannaschii* genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

One application of this embodiment is provided in Figure 4. Figure 4 provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium

storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, etc. A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software for reading the control logic and/or the data from the removable medium storage device 114 once inserted in the removable medium storage device 114.

A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and processing the genomic sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

Having generally described the invention, the same will be more readily understood by reference to the following examples, which are provided by way of illustration and are not intended as limiting.

Experimental

Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii

Example 1

A whole genome random sequencing method (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)) was used to obtain the complete genome sequence for *M. jannaschii*. A small insert plasmid library (2.5 Kbp average insert size) and a large insert lambda library (16 Kbp average insert size) were used as substrates for sequencing. The lambda library was used to form a genome scaffold and to verify the orientation and integrity of the contigs formed from the assembly of sequences from the plasmid library. All clones were sequenced from both ends to aid in ordering of contigs during the sequence assembly process. The average length of sequencing reads was 481 bp. A total of 36,718 sequences were assembled by means of the TIGR

Assembler (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995); Sutton G., *et al.*, *Genome Sci. Tech. 1*:9 (1995)). Sequence and physical gaps were closed using a combination of strategies (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)). The colinearity of the *in vivo* genome to the genome sequence was confirmed by comparing restriction fragments from six, rare cutter, restriction enzymes (Aat II, BamHI, Bgl II, Kpn I, Sma I, and Sst II) to those predicted from the sequence data. Additional confidence in the colinearity was provided by the genome scaffold produced by sequence pairs from 339 large-insert lambda clones, which covered 88% of the main chromosome. Open reading frames (ORFs) and predicted protein-coding regions were identified as described (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)) with some modification. In particular, the statistical prediction of *M. jannaschii* genes was performed with GeneMark (Borodovsky, M. & McIninch, J. *Comput. Chem. 17*:123 (1993)). Regular GeneMark uses nonhomogeneous Markov models derived from a training set of coding sequences and ordinary Markov models derived from a training set of noncoding sequences. Only a single 16S ribosomal RNA sequence of *M. jannaschii* was available in the public sequence databases before the whole genome sequence described here. Thus, the initial training set to determine parameters of a coding sequence Markov model was chosen as a set of ORFs >1000 nucleotides (nt). As an initial model for non-coding sequences, a zero-order Markov model with genome-specific nucleotide frequencies was used. The initial models were used at the first prediction step. The results of the first prediction were then used to compile a set of putative genes used at the second training step. Alternate rounds of training and predicting were continued until the set of predicted genes stabilized and the parameters of the final fourth-order model of coding sequences were derived. The regions predicted as noncoding were then used as a training set for a final model for noncoding regions. Cross-validation simulations demonstrated that the GeneMark program trained as described above was able to correctly identify coding regions of at least 96 nt in 94% of the cases and noncoding regions of the

same length in 96% of the cases. These values assume that the self-training method produced correct sequence annotation for compiled control sets. Comparison with the results obtained by searches against a nonredundant protein database (Fleischmann, R.D., *et al., Science 269*:496 (1995); Fraser, C.M., *et al., Science 270*:397 (1995)) demonstrated that almost all genes identified by sequence similarity were predicted by the GeneMark program as well. This observation provides additional confidence in genes predicted by GeneMark whose protein translations did not show significant similarity to known protein sequences. The predicted protein-coding regions were search against the Blocks database (Henikoff, S. & Henikoff, J.G., *Genomics 19*:97 (1994)] by means of BLIMPS (Wallace, J.C. & Henikoff, S., *CABIOS 8*:249 (1992)) to verify putative identifications and to identify potential functional motifs in predicted protein-coding regions that had no database match. Genes were assigned to known metabolic pathways. When a gene appeared to be missing from a pathway, the unassigned ORFs and the complete *M. jannaschii* genome sequence were searched with specific query sequences or motifs from the Blocks database. Hydrophobicity plots were performed on all predicted protein-coding regions by means of the Kyte-Doolittle algorithm (Kyte, J. & Doolittle, R.F., *J. Mol. Biol. 157*:105 (1982)) to identify potentially functionally relevant signatures in these sequences.

The *M. jannaschii* genome comprises three physically distinct elements: i) a large circular chromosome of 1,664,976 base pairs (bp) (SEQ ID NO:1), which contains 1682 predicted protein-coding regions and has a G+C content of 31.4%; ii) a large circular extrachromosomal element (ECE) (Zhao, H., *et al., Arch. Microbiol. 150*:178 (1988)) of 58,407 bp (SEQ ID NO:2), which contains 44 predicted protein coding regions and has a G+C content of 28.2%; and iii) a small circular ECE (Zhao, H., *et al., Arch. Microbiol. 150*:178 (1988)) of 16,550 bp (SEQ ID NO:3), which contains 12 predicted protein coding regions, and has a G+C content of 28.8%. With respect to its shape, size, G+C content, and gene density the main chromosome resembles that of *H. influenzae.* However, here the resemblance stops.

Of the 1743 predicted protein-coding regions reported previously for *H. influenzae*, 78% had a match in the public sequence database (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)). Of these, 58% were matches to genes with reasonably well defined function, while 20% were matches to genes whose function was undefined. Similar observations were made for the *M. genitalium* genome (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995); Fraser, C.M., *et al.*, *Science 270*:397 (1995)). Eighty-three percent of the predicted protein coding regions from *M. genitalium* have a counterpart in the *H. influenzae* genome. In contrast, only 38% of the predicted protein-coding regions from *M. jannaschii* match a gene in the database that could be assigned a putative cellular role with high confidence; 6% of the predicted protein-coding regions had matches to hypothetical proteins (Tables 2-3). Approximately 100 genes in *M. jannaschii* had marginal similarity to genes or segments of genes from the public sequence databases and could not be assigned a putative cellular role with high confidence. Only 11% of the predicted protein-coding regions from *H. influenzae* and 17% of the predicted protein coding regions from *M. genitalium* matched a predicted protein coding region from *M. jannaschii*. Clearly the *M. jannaschii* genome, and undoubtedly, therefore, all archaeal genomes are remarkably unique, as the phylogenetic position of these organisms would suggest.

Energy production in *M. jannaschii* occurs via the reduction of CO_2 with H_2 to produce methane. Genes for all of the known enzymes and enzyme complexes associated with methanogenesis (DiMarco, A.A., *et al.*, *Ann. Rev. Biochem. 59*:355 (1990)) were identified in *M. jannaschii* , the sequence and order of which are typical of methanogens. *M. jannaschii* appears to use both H_2 and formate as substrates for methanogenesis, but lacks the genes to use methanol or acetate. The ability to fix nitrogen has been demonstrated in a number of methanogens (Belay, N., *et al.*, *Nature 312*:286 (1984)) and all of the genes necessary for this pathway have been identified in *M. jannaschii* (Tables 2-3). In addition to its anabolic pathways, several scavenging molecules have been

identified in *M. jannaschii* that probably play a role in importing small organic compounds, such as amino acids, from the environment (Tables 2-3).

Three different pathways are known for the fixation of CO_2 into organic carbon: the non-cyclic, reductive acetyl-coenzyme A-carbon monoxide dehydrogenase pathway (Ljungdahl-Wood pathway), the reductive trichloroacetic acid (TCA) cycle, and the Calvin cycle. Methanogens fix carbon by the Ljungdahl-Wood pathway (Wood, H.G., *et al.*, *TIBS 11*:14 (1986)), which is facilitated by the carbon monoxide dehydrogenease enzyme complex (CODH) (Blaat, M., *Antonie van Leewenhoek 66*:187 (1994)). The complete Ljungdahl-Wood pathway, encoded in the *M. jannaschii* genome, depends on the methyl carbon in methanogenesis; however, methanogenesis can occur independently of carbon fixation.

Although genes encoding two enzymes required for gluconeogenesis (glucopyruvate oxidoreductase and phosphoenolpyruvate synthase) were found in the *M. jannaschii* genome, genes encoding other key intermediates of gluconeogenesis (fructose bisphosphatase and fructose 1,6-bisphosphate aldolase) were not been identified. Glucose catabolism by glycolysis also requires the aldolase, as well as phosphofructokinase, an enzyme that also was not found in *M. jannaschii* and has not been detected in any of the Archaea. In addition, genes specific for the Entner-Doudoroff pathway, an alternative pathway used by some microbes for the catabolism of glucose, were not identified in the genomic sequence. The presence of a number of nearly complete metabolic pathways suggests that some key genes are not recognizable at the sequence level, although we cannot exclude the possibility that *M. jannaschii* may use alternative metabolic pathways.

In general, *M. jannaschii* genes that encode proteins involved in the transport of small inorganic ions into the cell are homologs of bacterial genes. The genome includes many representatives of the ABC transporter family, as well as genes for exporting heavy metals (e.g., the chromate-resistance protein) and other toxic compounds (e.g., the norA drug efflux pump locus).

More than 20 predicted protein-coding regions have sequence similarity to polysaccharide biosynthetic enzymes. These genes have only bacterial homologs or are most closely related to their bacterial counterparts. The identified polysaccharide biosynthetic genes in *M. jannaschii* include those for the interconversion of sugars, activation of sugars to nucleotide sugars, and glycosyltransferases for the polymerization of nucleotide sugars into oligo- and polysaccharides that are subsequently incorporated into surface structures (Hartmann, E. and König, H., *Arch. Microbiol. 151*:274 (1989)). In an arrangement reminiscent of bacterial polysaccharide biosynthesis genes, many of the genes for *M. jannaschii* polysaccharide production are clustered together (Tables 2-3). The G+C content in this region is <95% of that in the rest of the *M. jannaschii* genome. A similar observation was made in *Salmonella typhimurium* (Jiang, X.M., *et al.*, *Mol. Microbiol. 5*:695 (1991)) in which the gene cluster for lipopolysaccharide O antigen has a significantly lower G+C ratio than the rest of the genome. In that case, the difference in G+C content was interpreted as meaning that the region originated by lateral transfer from another organism.

Of the three main multicomponent information processing systems (transcription, translation, and replication), translation appears the most universal in its overall makeup in that the basic translation machinery is similar in all three domains of life. *M. jannaschii* has two ribosomal RNA operons, designated A and B, and a separate 5S RNA gene that is associated with several transfer RNAs (tRNAs). Operon A has the organization, 16S - 23S - 5S, whereas operon B lacks the 5S component. An alanine tRNA is situated in the spacer region between the 16S and 23S subunits in both operons. The majority of proteins associated with the ribosomal subunits (especially the small subunit) are present in both Bacteria and Eukaryotes. However, the relatively protein-rich eukaryotic ribosome contains additional ribosomal proteins not found in the bacterial ribosome. A smaller number of bacteria-specific ribosomal proteins exist as well. The *M. jannaschii* genome contains all ribosomal proteins that are common to eukaryotes and bacteria. It shows no homologs of the bacterial-specific ribosomal proteins, but does possess homologs of a number of the eukaryotic-specific ones.

Homologs of all archaea-specific ribosomal proteins that have been reported to date (Lechner, K., *et al.*, *J. Mol. Evol. 29*:20 (1989); Köpke, A.K.E. and Wittmann-Liebold, B., *Can. J. Microbiol. 35*:11 (1989)) are found in *M. jannaschii*.

As previously shown for other archaea (Iwabe, N., *et al.*, *Proc. Natl. Acad. Sci. USA 86*:9355 (1989); Gogarten J.P., *et al.*, *Proc. Natl. Acad. Sci. USA 86*:6661 (1989); Brown, J.R. and Doolittle, W.F., *Proc. Natl. Acad. Sci. USA 92*:2441 (1995)), the *Methanococcus* translation elongation factors EF-1α (EF-Tu in bacteria) and EF-2 (EF-G in bacteria) are most similar to their eukaryotic counterparts. In addition, the *M. jannaschii* genome contains 11 translation initiation factor genes. Three of these genes encode the subunits homologous to those of the eukaryotic IF-2, and are reported here in the Archaea for the first time. A fourth initiation factor gene that encodes a second IF-2 is also found in *M. jannaschii*. This additional IF-2 gene is most closely related to the yeast protein FUN12 which, in turn, appears to be a homolog of the bacterial IF-2. It is not known which of the two IF-2-like initiation factors identified in *M. jannaschii* plays a role in directing the initiator tRNA to the start site of the mRNA. The fifth identified initiation factor gene in *M. jannaschii* encodes IF-1A, which has no bacterial homolog. The sixth gene encodes the hypusine-containing initiation factor eIF-5a. Two subunits of the translation initiation factor eIF-2B were identified in *M. jannaschii*. Finally, three putative adenososine 5′-triphosphate (ATP)-dependent helicases were identified that belong to the eIF-4a family of translation initiation factors.

Thirty-seven tRNA genes were identified in the *M. jannaschii* genome. Almost all amino acids encoded by two codons have a single tRNA, except for glutamic acid, which has two. Both an initiator and an internal methionyl tRNA are present. The two pyrimidine-ending isoleucine codons are covered by a single tRNA, while the third (AUA) seems covered by a related tRNA having a CAU anticodon. A single tRNA appears to cover the three isoleucine codons. Those amino acids encoded by four codons each have two tRNAs, one to cover the Y-, the other the R-ending, codons. Valine has a third tRNA, which is

specific for the GUG codon; and alanine has three tRNAs (two of which are in the spacer regions separating the 16S and 23S subunits in the two ribosomal RNA operons). Leucine, serine and arginine, all of which have six codons, each posses three corresponding tRNAs. The genes for the internal methionine and tryptophan tRNAs contain introns in the region of their anti-codon loops.

A tRNA also exists for selenocysteine (UGA codon). At least four genes in *M. jannaschii* contain internal stop codons that are potential selenocysteine codons: the α chain of formate dehydrogenase, coenzyme F420 reducing hydrognase, β-chain tungsten formyl methanofuran dehydrogenase, and a heterodisulfide reductase. Three genes with a putative role in selenocysteine metabolism were identified by their similarity to the *sel* genes from other organisms (Tables 2-3).

Recognizable homologs for four of the aminoacyl-tRNA synthetases (glutamine, asparagine, lysine, and cysteine) were not identified in the *M. jannaschii* genome. The absence of a glutaminyl-tRNA synthetase is not surprising in that a number of organisms, including at least one archaeon, have none (Wilcox, M., *Eur. J. Biochem. 11*:405 (1969); Martin, N.C., *et al., J. Mol. Biol. 101*:285 (1976); Martin, N.C., *et al., Biochemistry 16*:4672 (1977); Schon, A., *et al., Biochimie 70*:391 (1988); Soll, D. and RajBhandary, U., Eds. *Am. Soc. for Microbiol.* (1995)). In these instances, glutaminyl tRNA charging involves a post-charging conversion mechanism whereby the tRNA is charged by the glutamyl-tRNA synthetase with glutamic acid, which then is enzymatically converted to glutamine. A post-charging conversion is also involved in selenocysteine charging via the seryl-tRNA synthetase. A similar mechanism has been proposed for asparagine charging, but has never been demonstrated (Wilcox, M., *Eur. J. Biochem. 11*:405 (1969); Martin, N.C., *et al., J. Mol. Biol. 101*:285 (1976); Martin, N.C., *et al., Biochemistry 16*:4672 (1977); Schon, A., *et al., Biochimie 70*:391 (1988); Soll, D. and RajBhandary, U., Eds. *Am. Soc. for Microbiol.* (1995)). The inability to find homologs of the lysine and cysteine aminoacyl-tRNA synthetases is surprising because bacterial and eukaryotic versions in each instance show clear homology.

Aminoacyl-tRNA synthetases of *M. jannaschii* and other archaea resemble eukaryotic synthetases more closely than they resemble bacterial forms. The tryptophanyl synthetase is one of the more notable examples, because the *M. jannaschii* and eukaryotic version do not appear to be specifically related to the bacterial version (de Pouplana, R., *et al.*, *Proc. Natl. Acad. Sci., USA 93*:166 (1996)). Two versions of the glycyl synthetase are known in bacteria, one that is very unlike the version found in Archaea and Eukaryote and one that is an obvious homolog of it (Wagner, E.A., *et al.*, *J. Bacteriol. 177*:5179 (1995); Logan, D.T., *et al.*, *EMBO J. 14*:4156 (1995)).

Eleven genes encoding subunits of the DNA-dependent RNA polymerase were identified in the *M. jannaschii* genome. The sequence similarity between the subunits and their homologs in *Sulfolobus acidocaldarius* supports the evolutionary unity of the archaeal polymerase complex (Woese, C.R. and Wolfe, R.S., Eds. *The Bacteria, vol. VIII* (Academic Press, NY, 1985); Langer, D., *et al.*, *Proc. Natl. Acad. Sci. 92*:5768 (1995); Lanzendoerfer, M. *et al.*, *System. Appl. Microbiol. 16*:656 (1994)). All of the subunits found in *M. jannaschii* show greater similarity to their eukaryotic counterparts than to the bacterial homologs. The genes encoding the five largest subunits (A′, A″, B′, B″, D) have homologs in all organisms. Six genes encode subunits shared only by Archaea and Eukaryotes (E, H, K, L, and N). The *M. jannaschii* homolog of the *S. acidocaldarius* subunit E is split into two genes designated E′ and E″. *Sulfolobus acidocaldarius* also contains two additional small subunits of RNA polymerase, designated G and F, that have no counterparts in either Bacteria or Eukaryotes. No homolog of these subunits was identified in *M. jannaschii*.

The archaeal transcription initiation system is essentially the same as that found in Eukaryotes, and is radically different from the bacterial version (Klenk, H.P. and Doolittle, W.F., *Curr. Biol.* 4:920 (1994)). The central molecules in the former systems are the TATA-binding protein (TBP) and transcription factor B (TFIIB and TFIIIB in Eukaryotes, or simply TFB). In the eukaryotic systems, TBP and TFB are parts of larger complexes, and additional factors (such as

TFIIA and TFIIF) are used in the transcription process. However, the *M. jannaschii* genome does not contain obvious homologs of TFIIA and TFIIF.

Several components of the replication machinery were identified in *M jannaschii*. The *M. jannaschii* genome appears to encode a single DNA-dependent polymerase that is a member of the B family of polymerases (Bernard, A., *et al.*, *EMBO J. 6*:4219 (1987); Cullman, G., *et al.*, *Molec. Cell Biol. 15*:4661 (1995); Uemori, T., *et al.*, *J. Bacteriol. 117*:2164 (1995); Delarue, M., *et al.*, *Prot. Engineer. 3*:461 (1990); Gavin, K.A., *et al.*, *Science 270*:1667 (1995)). The polymerase shares sequence similarity and three motifs with other family B polymerases, including eukaryotic α, γ, and ϵ polymerases, bacterial polymerase II, and several archaeal polymerases. However, it is not homologous to bacterial polymerase I and has no homologs in *H. influenzae* or *M. genitalium*.

Primer recognition by the polymerase takes place through a structure-specific DNA binding complex, the replication factor complex (rfc) (Bernard, A., *et al.*, *EMBO J. 6*:4219 (1987); Cullman, G., *et al.*, *Molec. Cell Biol. 15*:4661 (1995); Uemori, T., *et al.*, *J. Bacteriol. 117*:2164 (1995); Delarue, M., *et al.*, *Prot. Engineer. 3*:461 (1990); Gavin, K.A., *et al.*, *Science 270*:1667 (1995)). In humans and yeast, the rfc is composed of five proteins: a large subunit and four small subunits that have an associated adenosine triphosphatase (ATPase) activity stimulated by proliferating cell nuclear antigen (PCNA). Two genes in *M. jannaschii* are putative members of a eukaryotic-like replication factor complex. One of the genes in *M. jannaschii* is a putative homolog of the large subunit of the rfc, whereas the second is a putative homolog of one of the small subunits. Among Eukaryotes, the rfc proteins share sequence similarity in eight signature domains (Bernard, A., *et al.*, *EMBO J. 6*:4219 (1987); Cullman, G., *et al.*, *Molec. Cell Biol. 15*:4661 (1995); Uemori, T., *et al.*, *J. Bacteriol. 117*:2164 (1995); Delarue, M., *et al.*, *Prot. Engineer. 3*:461 (1990); Gavin, K.A., *et al.*, *Science 270*:1667 (1995)). Domain I is conserved only in the large subunit among Eukaryotes and is similar in sequence to DNA ligases. This domain is missing in the large-subunit homolog in *M. jannaschii*. The remaining domains in the two *M. jannaschii* genes are well-conserved relative to the eukaryotic homologs. Two

features of the sequence similarity in these domains are of particular interest. First, domain II (an ATPase domain) of the small-subunit homolog is split between two highly conserved amino acids (lysine and threonine) by an intervening sequence of unknown function. Second, the sequence of domain VI has regions that are useful for distinguishing between bacterial and eukaryotic rfc proteins (Bernard, A., *et al.*, *EMBO J. 6*:4219 (1987); Cullman, G., *et al.*, *Molec. Cell Biol. 15*:4661 (1995); Uemori, T., *et al.*, *J. Bacteriol. 117*:2164 (1995); Delarue, M., *et al.*, *Prot. Engineer. 3*:461 (1990); Gavin, K.A., *et al.*, *Science 270*:1667 (1995)); the rfc sequence for *M. jannaschii* shares the characteristic eukaryotic signature in this domain.

We have attempted to identify an origin of replication by searching the *M. jannaschii* genome sequence with a variety of bacterial and eukaryotic replication-origin consensus sequences. Searches with oriC, ColE1, and autonomously replicating sequences from yeast (Bernard, A., *et al.*, *EMBO J. 6*:4219 (1987); Cullman, G., *et al.*, *Molec. Cell Biol. 15*:4661 (1995); Uemori, T., *et al.*, *J. Bacteriol. 117*:2164 (1995); Delarue, M., *et al.*, *Prot. Engineer. 3*:461 (1990); Gavin, K.A., *et al.*, *Science 270*:1667 (1995)) did not identify an origin of replication. With respect to the related cellular processes of replication initiation and cell division, the *M. jannaschii* genome contains two genes that are putative homologs of Cdc54, a yeast protein that belongs to a family of putative DNA replication initiation proteins (Whitbred, L.A. and Dalton, S., *Gene 155*:113 (1995)). A third potential regulator of cell division in *M. jannaschii* is 55% similar at the amino acid level to *pelota*, a *Drosophila* protein involved in the regulation of the early phases of meiotic and mitotic cell division (Eberhart, C.G. and Wasserman, S.A., *Development 121*:3477 (1995)).

In contrast to the putative rfc complex and the initiation of DNA replication, the cell division proteins from *M. jannaschii* most resemble their bacterial counterparts (Rothfield, L.I. and Zhao, C.R., *Cell 84*:183 (1996); Lutkenhaus, J., *Curr. Opp. Gen. Devel. 3*:783 (1993)). Two genes similar to that encoding FtsZ, a ubiquitous bacterial protein, are found in *M. jannaschii*. FtsZ

is a polymer-forming, guanosine triphosphate (GTP)-hydrolyzing protein with tubulin-like elements; it is localized to the site of septation and forms a constricting ring between the dividing cells. One gene similar to FtsJ, a bacterial cell division protein of undetermined function, also is found in *M. jannaschii*. Three additional genes (MinC, MinD, and MinE) function in concert in Bacteria to determine the site of septation during cell division. In *M. jannaschii*, three MinD-like genes were identified, but none for MinC or MinE. Neither spindle-associated proteins characteristic of eukaryotic cell division nor bacterial mechanochemical enzymes necessary for partitioning the condensed chromosomes were detected in the *M. jannaschii* genome. Taken together, these observations raise the possibility that cell division in *M. jannaschii* might occur via a mechanism specific for the Archaea.

The structural and functional conservation of the signal peptide of secreted proteins in Archaea, Bacteria, and Eukaryotes suggests that the basic mechanisms of membrane targeting and translocation may be similar among all three domains of life. The secretory machinery of *M. jannaschii* appears a rudimentary apparatus relative to that of bacterial and eukaryotic systems and consists of (i) a signal peptidase (SP) that cleaves the signal peptide of translocating proteins, (ii) a preprotein translocase that is the major constituent of the membrane-localized translocation channel, (iii) a ribonucleoprotein complex (signal recognition particle, SRP) that binds to the signal peptide and guides nascent proteins to the cell membrane, and (iv) a docking protein that acts as a receptor for the SRP. The 7S RNA component of the SRP from *M. jannaschii* shows a highly conserved structural domain shared by other Archaea, Bacteria, and Eukaryotes (Kaine, B.P. and Merkel, V.L., *J. Bacteriol. 171*:4261 (1989); Poritz, M.A. *et al., Cell 55*:4 (1988)). However, the predicted secondary structure of the 7S RNA SRP component in Archaea is more like that found in Eukaryotes than in Bacteria (Kaine, B.P. and Merkel, V.L., *J. Bacteriol. 171*:4261 (1989); Poritz, M.A. *et al., Cell 55*:4 (1988)). The SP and docking proteins from *M. jannaschii* are most similar to their eukaryotic counterparts; the translocase is most similar to the SecY translocation-associated protein in *Escherichia coli*.

A second distinct signal peptide is found in the flagellin genes of *M. jannaschii*. Alignment of flagellin genes from *M. voltae* (Faguy, D.M., *et al.*, *Can. J. Microbiol. 40*:67 (1994); Kalmokoff, M.L., *et al.*, *Arch. Microbiol. 157*:481 (1992)) and *M. jannaschii* reveals a highly conserved NH_2-terminus (31 of the first 50 residues are identical in all of the mature flagellins). The peptide sequence of the *M. jannaschii* flagellin indicates that the protein is cleaved after the canonical Gly-12 position, and it is proposed to be similar to type-IV pilins of Bacteria (Faguy, D.M., *et al.*, *Can. J. Microbiol. 40*:67 (1994); Kalmokoff, M.L., *et al.*, *Arch. Microbiol. 157*:481 (1992)).

Five histone genes are present in the *M. jannaschii* genome--three on the main chromosome and two on the large ECE. These genes are homologs of eukaryotic histones (H2a, H2b, H3, and H4) and of the eukaryotic transcription-related CAAT-binding factor CBF-A (Sandman, K., *et al.*, *Proc. Natl. Acad. Sci. USA 87*:5788 (1990)). The similarity between archaeal and eukaryotic histones suggests that the two groups of organisms resemble one another in the roles histones play both in genome supercoiling dynamics and in gene expression. The five *M. jannaschii* histone genes show greatest similarity among themselves even though a histone sequence is available from the closely related species, *Methanococcus voltae*. This intraspecific similarity suggests that the gene duplications that produced the five histone genes occurred on the *M. jannaschii* lineage per se.

Self-splicing portions of a peptide sequence that generally encode a DNA endonuclease activity are called inteins, in analogy to introns (Kane, P.M., *et al.*, *Science 250*:651 (1990); Hirata, R., *et al.*, *J. Biol. Chem. 265*:6726 (1990); Cooper, A. and Stevens, T., *TIBS 20*:351 (1995); Xu, M.Q., *et al.*, *Cell 75*:1371 (1993); Perler *et al.*, *Proc. Natl. Acad. Sci. USA 89*:5577 (1992); Cooper *et al.*, *EMBO J. 12*:2575 (1993); Michel *et al.*, *Biochimie 64*:867 (1992); Pietrokovski S., *Prot. Sci. 3*:2340 (1994). Most inteins in the *M. jannaschii* genome were identified by (i) similarity of the bounding exteins to other proteins, (ii) similarity of the inteins to those previously described, (iii) presence of the dodecapeptide endonuclease motifs, and (iv) canonical intein-extein junction sequences. In two

instances (MJ0832 and MJ0043), the similarity to other database sequences did not unambiguously define the NH_2-terminal extein-intein junction, so it was necessary to rely on consensus sequences to select the putative site. The inteins in MJ1042 and MJ0542 have previously uncharacterized COOH-terminal splice junctions, GNC and FNC, respectively).

The sequences remaining after an intein is excised are called exteins, in analogy to exons. Exteins are spliced together after the excision of one or more inteins to form functional proteins. The biological significance and role of inteins are not clearly understood (Kane, P.M., *et al.*, *Science 250*:651 (1990); Hirata, R., *et al.*, *J. Biol. Chem. 265*:6726 (1990); Cooper, A. and Stevens, T., *TIBS 20*:351 (1995); Xu, M.Q., *et al.*, *Cell 75*:1371 (1993); Perler *et al.*, *Proc. Natl. Acad. Sci. USA 89*:5577 (1992); Cooper *et al.*, *EMBO J. 12*:2575 (1993); Michel *et al.*, *Biochimie 64*:867 (1992); Pietrokovski S., *Prot. Sci. 3*:2340 (1994)). Fourteen genes in the *M. jannaschii* genome contain 18 putative inteins, a significant increase in the approximately 10 intein-containing genes that have been described (Kane, P.M., *et al.*, *Science 250*:651 (1990); Hirata, R., *et al.*, *J. Biol. Chem. 265*:6726 (1990); Cooper, A. and Stevens, T., *TIBS 20*:351 (1995); Xu, M.Q., *et al.*, *Cell 75*:1371 (1993); Perler *et al.*, *Proc. Natl. Acad. Sci. USA 89*:5577 (1992); Cooper *et al.*, *EMBO J. 12*:2575 (1993); Michel *et al.*, *Biochimie 64*:867 (1992); Pietrokovski S., *Prot. Sci. 3*:2340 (1994)) (Table 4). The only previously described inteins in the Archaea are in the DNA polymerase genes of the Thermococcales (Kane, P.M., *et al.*, *Science 250*:651 (1990); Hirata, R., *et al.*, *J. Biol. Chem. 265*:6726 (1990); Cooper, A. and Stevens, T., *TIBS 20*:351 (1995); Xu, M.Q., *et al.*, *Cell 75*:1371 (1993); Perler *et al.*, *Proc. Natl. Acad. Sci. USA 89*:5577 (1992); Cooper *et al.*, *EMBO J. 12*:2575 (1993); Michel *et al.*, *Biochimie 64*:867 (1992); Pietrokovski S., *Prot. Sci. 3*:2340 (1994)). The *M. jannaschii* DNA polymerase gene has two inteins in the same locations as those in *Pyrococcus* sp. strain KOD1. In this case, the exteins exhibit 46% amino acid identity, whereas intein 2 of the two organisms has only 33% identity. This divergence suggests that intein 2 has not been recently (laterally) transferred between the Thermococcales and *M. jannaschii*. In contrast, the intein 1

sequences are 56% identical, more than that of the gene containing them, and comparable to the divergence of inteins within the Thermococcales. This high degree of sequence similarity might be the result of an intein transfer more recent than the splitting of these species. The large number of inteins found in *M. jannaschii* led us to question whether these inteins have been increasing in number by moving within the genome. If this were so, we would expect to find some pairs of inteins that are particularly similar. Comparisons of these and other available intein sequences showed that the closest relationships are those noted above linking the DNA polymerase inteins to correspondingly positioned elements in the Thermococcales. Within *M. jannaschii,* the highest identity observed was 33% for a 380-bp portion of two inteins. This finding suggests that the diversification of the inteins predates the divergence of the *M. jannaschii* and *Pyrococcus* DNA polymerases.

Three families of repeated genetic elements were identified in the *M. jannaschii* genome. Within two of the families, at least two members were identified as ORFs with a limited degree of sequence similarity to bacterial transposases. Members of the first family, designated *ISAMJ1*, are repeated 10 times on the main chromosome and once on the large ECE (Fig. 2). There is no sequence similarity between the IS elements in *M. jannaschii* and the IS*M1* mobile element described previously for *Methanobrevibacter smithii* (Hamilton, P.T. *et al., Mol. Gen. Genet. 200*:47 (1985)). Two members of this family were identified as ORFs and are 27% identical (at the amino acid sequence level) to a transposase from *Bacillus thuringiensis* (IS*240*; GenBank accession number M23741). Relative to these two members, the remaining members of the IS*AMJ1* family are missing an internal region of several hundred nucleotides (Fig. 2). With one exception, all members of this family end with 16-bp terminal inverted repeats typical of insertion sequences. One member is missing the terminal repeat at its 5′ end. The second family consists of two ORFs that are identical across 928 bp. The ORFs are 23% identical at the amino acid sequence level to the COOH-terminus of a transposase from *Lactococcus lactis* (IS*982*; GenBank

accession number L34754). Neither of the members of the second family contains terminal inverted repeats.

Eighteen copies of the third family of repeated genetic structures (Fig. 3) are distributed fairly evenly around the *M. jannaschii* genome. Unlike the genetic elements described above, none of the components of this repeat unit appears to have coding potential. The repeat structure is composed of a long segment followed by one to 25 tandem repetitions of a short segment. The short segments are separated by sequence that is unique within and among the complete repeat structure. Three similar types of short segments were identified; however, the type of short repeat is consistent within each repeat structure, except for variation of the last short segment in six repeat structures. Similar tandem repeats of short segments have been observed in Bacteria and other Archaea (Mojica, F.J.M., *et al.*, *Mol. Micro. 17*:85 (1995)) and have been hypothesized to participate in chromosome partitioning during cell division.

The 16-kbp ECE from *M. jannaschii* contains 12 ORFs, none of which had a significant full-length match to any published sequence. The 58-kbp ECE contains 44 predicted protein-coding regions, 5 of which had matches to genes in the database. Two of the genes are putative archaeal histones, one is a sporulation-related protein (SOJ protein), and two are type I restriction modification enzymes. There are several instances in which predicted protein-coding regions or repeated genetic elements on the large ECE have similar counterparts on the main chromosome of *M. jannaschii*. The degree of nucleotide sequence similarity between genes present on both the ECE and the main chromosome ranges from 70 to 90%, suggesting that there has been relatively recent exchange of at least some genetic material between the large ECE and the main chromosome.

All the predicted protein-coding regions from *M. jannaschii* were searched against each other in order to identify families of paralogous genes (genes related by gene duplication, not speciation). The initial criterion for grouping paralogs was >30% amino acid sequence identity over 50 consecutive amino acid residues. Groups of predicted protein-coding regions were then

aligned and inspected individually to ensure that the sequence similarity extended over most of their lengths. This curatorial process resulted in the identification of more than 100 gene families, half of which have no database matches. The largest identified gene family (16 members: MJ0625, MJECL28, MJ1076, MJ1006, MJ1659, MJ0075, MJ1609, MJECL19, MJECL18, MJ0147, MJ0801, MJ1301, MJ0632, MJ1010, MJ0074, and MJ0439) contains almost 1% of the total predicted protein-coding regions in *M. jannaschii*.

Despite the availability for comparison of two complete bacterial genomes and several hundred megabase pairs of eukaryotic sequence data, the majority of genes in *M. jannaschii* cannot be identified on the basis of sequence similarity. Previous evidence for the shared common ancestry of the Archaeal and Eukaryotic was based on a small set gene sequences (Iwabe, N., *et al., Proc. Natl. Acad. Sci. USA 86*:9355 (1989); Gogarten J.P., *et al., Proc. Natl. Acad. Sci. USA 86*:6661 (1989); Brown, J.R. and Doolittle, W.F., *Proc. Natl. Acad. Sci. USA 92*:2441 (1995)). The complete genome of *M. jannaschii* allows us to move beyond a "gene by gene" approach to one that encompasses the larger picture of metabolic capacity and cellular systems. The anabolic genes of *M. jannaschii* (especially those related to energy production and nitrogen fixation) reveal an ancient metabolic world shared largely by Bacteria and Archaea. That many basic autotrophic pathways appear to have a common evolutionary origin suggests that the most recent universal common ancestor to all three domains of extant life had the capacity for autotrophy. The Archaea and Bacteria also share structural and organizational features that the most recent universal prokaryotic ancestors also likely possessed, such as circular genomes and genes organized as operons. In contrast, the cellular information-processing and secretion systems in *M. jannaschii* demonstrate the common ancestry of Eukaryotes and Archaea. Although there are components of these systems are present in all three domains, their apparent refinement over time–especially transcription and translation–indicate that the Archaea and Eukaryotes share a common evolutionary trajectory independent of the lineage of Bacteria.

Example 2

Preparation of PCR Primers and Amplification of DNA

Various fragments of the *Methanococcus jannaschii* genome, such as those disclosed in Tables 2(a), 2(b) and 3 can be used, in accordance with the present invention, to prepare PCR primers. The PCR primers are preferably at least 15 bases, and more preferably at least 18 bases in length. When selecting a primer sequence, it is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. The PCR primers are useful during PCR cloning of the ORFs described herein.

Example 3

Gene expression from DNA Sequences Corresponding to ORFs

A fragment of the *Methanococcus jannaschii* genome (preferably, a protein-encoding sequence) provided in Tables 2(a), 2(b) or 3 is introduced into an expression vector using conventional technology (techniques to transfer cloned sequences into expression vectors that direct protein translation in mammalian, yeast, insect or bacterial expression systems are well known in the art). Commercially available vectors and expression systems are available from a variety of suppliers including Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence may be optimized for the particular expression organism, as explained by Hatfield *et al.*, U.S. Pat. No. 5,082,767, which is hereby incorporated by reference.

The following is provided as one exemplary method to generate polypeptide(s) from a cloned ORF of the *Methanococcus* genome whose sequence is provided in SEQ ID NOS: 1, 2 and 3. A poly A sequence can be

added to the construct by, for example, splicing out the poly A sequence from pSG5 (Stratagene) using *Bgl*I and *Sal*I restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene) for use in eukaryotic expression systems. pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex thymidine kinase promoter and the selectable neomycin gene. The *Methanococcus* DNA is obtained by PCR from the bacterial vector using oligonucleotide primers complementary to the *Methanococcus* DNA and containing restriction endonuclease sequences for *Pst*I incorporated into the 5′ primer and *Bgl*II at the 5′ end of the corresponding *Methanococcus* DNA 3′ primer, taking care to ensure that the *Methanococcus* DNA is positioned such that its followed with the poly A sequence. The purified fragment obtained from the resulting PCR reaction is digested with *Pst*I, blunt ended with an exonuclease, digested with *Bgl*II, purified and ligated to pXT1, now containing a poly A sequence and digested *Bgl*II.

The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Technologies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600 ug/ml G418 (Sigma, St. Louis, Missouri). The protein is preferably released into the supernatant. However if the protein has membrane binding domains, the protein may additionally be retained within the cell or expression may be restricted to the cell surface.

Since it may be necessary to purify and locate the transfected product, synthetic 15-mer peptides synthesized from the predicted *Methanococcus* DNA sequence are injected into mice to generate antibody to the polypeptide encoded by the *Methanococcus* DNA.

If antibody production is not possible, the *Methanococcus* DNA sequence is additionally incorporated into eukaryotic expression vectors and expressed as a chimeric with, for example, ß-globin. Antibody to ß-globin is used to purify the chimeric. Corresponding protease cleavage sites engineered between the ß-globin

gene and the *Methanococcus* DNA are then used to separate the two polypeptide fragments from one another after translation. One useful expression vector for generating ß-globin chimerics is pSG5 (Stratagene). This vector encodes rabbit ß-globin. Intron II of the rabbit ß-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques as described are well known to those skilled in the art of molecular biology. Standard methods are available from the technical assistance representatives from Stratagene, Life Technologies, Inc., or Promega. Polypeptides may additionally be produced from either construct using in vitro translation systems such as In vitro Express™ Translation Kit (Stratagene).

Example 4

E. coli Expression of a *M. jannaschii* ORF and protein purification

A *M. jannaschii* ORF described in Table 2(a), 2(b), or 3 is selected and amplified using PCR oligonucleotide primers designed from the nucleotide sequences flanking the selected ORF and/or from portions of the ORF's NH_2- or COOH-terminus. Additional nucleotides containing restriction sites to facilitate cloning are added to the 5′ and 3′ sequences, respectively.

The restriction sites are selected to be convenient to restriction sites in the bacterial expression vector pD10 (pQE9), which is used for bacterial expression. (Qiagen, Inc. 9259 Eton Avenue, Chatsworth, CA, 91311). [pD10]pQE9 encodes ampicillin antibiotic resistance ("Amp'") and contains a bacterial origin of replication ("ori"), an IPTG inducible promoter, a ribosome binding site ("RBS"), a 6-His tag and restriction enzyme sites.

The amplified *M. jannaschii* DNA and the vector pQE9 both are digested with SalI and XbaI and the digested DNAs are then ligated together. Insertion of the *M. jannaschii* DNA into the restricted pQE9 vector places the *M. jannaschii* coding region downstream of and operably linked to the vector's IPTG-inducible

promoter and in-frame with an initiating AUG appropriately positioned for translation of the *M. jannaschii* protein.

The ligation mixture is transformed into competent *E. coli* cells using standard procedures. Such procedures are described in Sambrook *et al.*, Molecular Cloning: a Laboratory Manual, 2nd Ed.; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1989). *E. coli* strain M15/rep4, containing multiple copies of the plasmid pREP4, which expresses lac repressor and confers kanamycin resistance ("Kan"), is used in carrying out the illustrative example described herein. This strain, which is only one of many that are suitable for expressing *M. jannaschii* protein, is available commercially from Qiagen.

Transformants are identified by their ability to grow on LB plates in the presence of ampicillin and kanamycin. Plasmid DNA is isolated from resistant colonies and the identity of the cloned DNA confirmed by restriction analysis. Clones containing the desired constructs are grown overnight ("O/N") in liquid culture in LB media supplemented with both ampicillin (100 μg/ml) and kanamycin (25 μg/ml).

The O/N culture is used to inoculate a large culture, at a dilution of approximately 1:100 to 1:250. The cells are grown to an optical density at 600nm ("OD600") of between 0.4 and 0.6. Isopropyl-B-D-thiogalactopyranoside ("IPTG") is then added to a final concentration of 1 mM to induce transcription from *lac* repressor sensitive promoters, by inactivating the *lacI* repressor. Cells subsequently are incubated further for 3 to 4 hours. Cells then are harvested by centrifugation and disrupted, by standard methods. Inclusion bodies are purified from the disrupted cells using routine collection techniques, and protein is solubilized from the inclusion bodies into 8M urea. The 8M urea solution containing the solubilized protein is passed over a PD-10 column in 2X phosphate-buffered saline ("PBS"), thereby removing the urea, exchanging the buffer and refolding the protein. The protein is purified by a further step of chromatography to remove endotoxin followed by sterile filtration. The sterile filtered protein preparation is stored in 2X PBS at a concentration of 95 μ/ml.

Example 5

Cloning and Expression of a M. jannaschii protein in a Baculovirus Expression System

A *M. jannaschii* ORF described in Table 2(a), 2(b), or 3 is selected and amplified as above. The amplified DNA is isolated from a 1% agarose gel using a commercially available kit ("Geneclean," BIO 101 Inc., La Jolla, Ca.). The DNA then is digested with XbaI and again purified on a 1% agarose gel. This DNA is designated herein as F2.

The vector pA2-GP is used to express the *M. jannaschii* protein in the baculovirus expression system as described in Summers *et al.*, A Manual of Methods for Baculovirus Vectors and Insect Cell Culture Procedures, Texas Agricultural Experimental Station Bulletin No. 1555 (1987). The pA2-GP expression vector contains the strong polyhedrin promoter of the *Autographa californica* nuclear polyhedrosis virus (AcMNPV) followed by convenient restriction sites. The signal peptide of AcMNPV gp67, including the N-terminal methionine, is located just upstream of a BamHI site. The polyadenylation site from the simian virus 40 ("SV40") is used for efficient polyadenylation. For an easy selection of recombinant virus, the beta-galactosidase gene from *E. coli* is inserted in the same orientation as the polyhedrin promoter and is followed by the polyadenylation signal of the polyhedrin gene. The polyhedrin sequences are flanked at both sides by viral sequences for cell-mediated homologous recombination with wild-type viral DNA to generate viable virus that express the cloned polynucleotide.

Many other baculovirus vectors could be used in place of pA2-GP, such as pAc373, pVL941 and pAcIM1 provided, as those of skill readily will appreciate, that construction provides appropriately located signals for transcription, translation, trafficking and the like, such as an in-frame AUG and a signal peptide, as required. Such vectors are described in Luckow *et al.*, *Virology 170*: 31-39, among others.

The plasmid is digested with the restriction enzyme XbaI and then is dephosphorylated using calf intestinal phosphatase, using routine procedures known in the art. The DNA is then isolated from a 1% agarose gel using a commercially available kit ("Geneclean" BIO 101 Inc., La Jolla, Ca.). This vector DNA is designated herein "V".

Fragment F2 and the dephosphorylated plasmid V2 are ligated together with T4 DNA ligase. *E. coli* HB101 cells are transformed with ligation mix and spread on culture plates. Bacteria are identified that contain the plasmid with the *M. jannaschii* gene by digesting DNA from individual colonies using XbaI and then analyzing the digestion product by gel electrophoresis. The sequence of the cloned fragment is confirmed by DNA sequencing. This plasmid is designated herein pBac*M. jannaschii*.

5 µg of the plasmid pBac*M. jannaschii* is co-transfected with 1.0 µg of a commercially available linearized baculovirus DNA ("BaculoGold™ baculovirus DNA", Pharmingen, San Diego, CA.), using the lipofection method described by Felgner *et al.*, Proc. Natl. Acad. Sci. USA 84: 7413-7417 (1987). 1µg of BaculoGold™ virus DNA and 5 µg of the plasmid pBac*M. jannaschii* are mixed in a sterile well of a microtiter plate containing 50 µl of serum-free Grace's medium (Life Technologies Inc., Gaithersburg, MD). Afterwards 10 µl Lipofectin plus 90 µl Grace's medium are added, mixed and incubated for 15 minutes at room temperature. Then the transfection mixture is added drop-wise to Sf9 insect cells (ATCC CRL 1711) seeded in a 35 mm tissue culture plate with 1 ml Grace's medium without serum. The plate is rocked back and forth to mix the newly added solution. The plate is then incubated for 5 hours at 27°C. After 5 hours the transfection solution is removed from the plate and 1 ml of Grace's insect medium supplemented with 10% fetal calf serum is added. The plate is put back into an incubator and cultivation is continued at 27°C for four days.

After four days the supernatant is collected and a plaque assay is performed, as described by Summers and Smith, cited above. An agarose gel with "Blue Gal" (Life Technologies Inc., Gaithersburg) is used to allow easy identification and isolation of gal-expressing clones, which produce blue-stained

plaques. (A detailed description of a "plaque assay" of this type can also be found in the user's guide for insect cell culture and baculovirology distributed by Life Technologies Inc., Gaithersburg, page 9-10).

Four days after serial dilution, the virus is added to the cells. After appropriate incubation, blue stained plaques are picked with the tip of an Eppendorf pipette. The agar containing the recombinant viruses is then resuspended in an Eppendorf tube containing 200 µl of Grace's medium. The agar is removed by a brief centrifugation and the supernatant containing the recombinant baculovirus is used to infect Sf9 cells seeded in 35 mm dishes. Four days later the supernatants of these culture dishes are harvested and then they are stored at 4°C. A clone containing properly inserted hESSB I, II and III is identified by DNA analysis including restriction mapping and sequencing. This is designated herein as V-*M. jannaschii.*

Sf9 cells are grown in Grace's medium supplemented with 10% heat-inactivated FBS. The cells are infected with the recombinant baculovirus V-*M. jannaschii* at a multiplicity of infection ("MOI") of about 2 (about 1 to about 3). Six hours later the medium is removed and is replaced with SF900 II medium minus methionine and cysteine (available from Life Technologies Inc., Gaithersburg). 42 hours later, 5 µCi of ^{35}S-methionine and 5 µCi ^{35}S-cysteine (available from Amersham) are added. The cells are further incubated for 16 hours and then they are harvested by centrifugation, lysed and the labeled proteins are visualized by SDS-PAGE and autoradiography.

Example 6

Cloning and Expression in Mammalian Cells

Most of the vectors used for the transient expression of a *M. jannaschii* gene in mammalian cells should carry the SV40 origin of replication. This allows the replication of the vector to high copy numbers in cells (e.g., COS cells) which

express the T antigen required for the initiation of viral DNA synthesis. Any other mammalian cell line can also be utilized for this purpose.

A typical mammalian expression vector contains the promoter element, which mediates the initiation of transcription of mRNA, the protein-coding sequence, and signals required for the termination of trancription and polyadenylation of the transcript. Additional elements include enhancers, Kozak sequences and intervening sequences flanked by donor and acceptor sites for RNA splicing. Highly efficient transcription can be achieved with the early and late promoters from SV40, the long terminal repeats (LTRs) from Retroviruses, e.g., RSV, HTLVI, HIVI and the early promoter of the cytomegalovirus (CMV). However, cellular signals can also be used (e.g., human actin promoter). Suitable expression vectors for use in practicing the present invention include, for example, vectors such as pSVL and pMSG (Pharmacia, Uppsala, Sweden), pRSVcat (ATCC 37152), pSV2dhfr (ATCC 37146) and pBC12MI (ATCC 67109). Mammalian host cells that could be used include, human Hela, 283, H9 and Jurkart cells, mouse NIH3T3 and C127 cells, Cos 1, Cos 7 and CV1, African green monkey cells, quail QC1-3 cells, mouse L cells and Chinese hamster ovary cells.

Alternatively, the gene can be expressed in stable cell lines that contain the gene integrated into a chromosome. The co-transfection with a selectable marker such as dhfr, gpt, neomycin, hygromycin allows the identification and isolation of the transfected cells.

The transfected gene can also be amplified to express large amounts of the encoded protein. The DHFR (dihydrofolate reductase) is a useful marker to develop cell lines that carry several hundred or even several thousand copies of the gene of interest. Another useful selection marker is the enzyme glutamine synthase (GS) (Murphy *et al.*, *Biochem J. 227*:277-279 (1991); Bebbington *et al.*, *Bio/Technology 10*:169-175 (1992)). Using these markers, the mammalian cells are grown in selective medium and the cells with the highest resistance are selected. These cell lines contain the amplified gene(s) integrated into a

chromosome. Chinese hamster ovary (CHO) cells are often used for the production of proteins.

The expression vectors pC1 and pC4 contain the strong promoter (LTR) of the Rous Sarcoma Virus (Cullen *et al.*, *Molecular and Cellular Biology*, 438-447 (March, 1985)) plus a fragment of the CMV-enhancer (Boshart *et al.*, *Cell 41*:521-530 (1985)). Multiple cloning sites, e.g., with the restriction enzyme cleavage sites BamHI, XbaI and Asp718, facilitate the cloning of the gene of interest. The vectors contain in addition the 3′ intron, the polyadenylation and termination signal of the rat preproinsulin gene.

Example 6(a): Cloning and Expression in COS Cells

The expression plasmid, p*M. jannaschii* HA, is made by cloning a cDNA encoding a *M. jannaschii* protein into the expression vector pcDNAI/Amp (which can be obtained from Invitrogen, Inc.).

The expression vector pcDNAI/amp contains: (1) an *E. coli* origin of replication effective for propagation in *E. coli* and other prokaryotic cells; (2) an ampicillin resistance gene for selection of plasmid-containing prokaryotic cells; (3) an SV40 origin of replication for propagation in eukaryotic cells; (4) a CMV promoter, a polylinker, an SV40 intron, and a polyadenylation signal arranged so that a cDNA conveniently can be placed under expression control of the CMV promoter and operably linked to the SV40 intron and the polyadenylation signal by means of restriction sites in the polylinker.

A DNA fragment encoding the *M. jannaschii* protein and an HA tag fused in frame to its 3′ end is cloned into the polylinker region of the vector so that recombinant protein expression is directed by the CMV promoter. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein described by Wilson *et al.*, *Cell 37*:767 (1984). The fusion of the HA tag to the target protein allows easy detection of the recombinant protein with an antibody that recognizes the HA epitope.

The PCR amplified DNA fragment (generated as described above) and the vector, pcDNAI/Amp, are digested with HindIII and XhoI and then ligated. The ligation mixture is transformed into *E. coli* strain SURE (available from Stratagene Cloning Systems, 11099 North Torrey Pines Road, La Jolla, CA 92037), and the transformed culture is plated on ampicillin media plates which then are incubated to allow growth of ampicillin resistant colonies. Plasmid DNA is isolated from resistant colonies and examined by restriction analysis and gel sizing for the presence of the *M. jannaschii* protein-encoding fragment.

For expression of recombinant *M. jannaschii*, COS cells are transfected with an expression vector, as described above, using DEAE-DEXTRAN, as described, for instance, in Sambrook *et al.*, Molecular Cloning: a Laboratory Manual, Cold Spring Laboratory Press, Cold Spring Harbor, New York (1989). Cells are incubated under conditions for expression of *M. jannaschii* protein by the vector.

Expression of the *M. jannaschii* HA fusion protein is detected by radiolabelling and immunoprecipitation, using methods described in, for example Harlow *et al.*, Antibodies: A Laboratory Manual, 2nd Ed.; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1988). To this end, two days after transfection, the cells are labeled by incubation in media containing ^{35}S-cysteine for 8 hours. The cells and the media are collected, and the cells are washed and the lysed with detergent-containing RIPA buffer: 150 mM NaCl, 1% NP-40, 0.1% SDS, 1% NP-40, 0.5% DOC, 50 mM TRIS, pH 7.5, as described by Wilson *et al.* cited above. Proteins are precipitated from the cell lysate and from the culture media using an HA-specific monoclonal antibody. The precipitated proteins then are analyzed by SDS-PAGE gels and autoradiography. An expression product of the expected size is seen in the cell lysate, which is not seen in negative controls.

Example 6(b): Cloning and Expression in CHO Cells

The vector pC1 is used for the expression of a *M. jannaschii* protein. Plasmid pC1 is a derivative of the plasmid pSV2-dhfr [ATCC Accession No. 37146]. Both plasmids contain the mouse DHFR gene under control of the SV40 early promoter. Chinese hamster ovary- or other cells lacking dihydrofolate activity that are transfected with these plasmids can be selected by growing the cells in a selective medium (alpha minus MEM, Life Technologies) supplemented with the chemotherapeutic agent methotrexate. The amplification of the DHFR genes in cells resistant to methotrexate (MTX) has been well documented (see, e.g., Alt, F.W., Kellems, R.M., Bertino, J.R., and Schimke, R.T., 1978, J. Biol. Chem. 253:1357-1370, Hamlin, J.L. and Ma, C. 1990, Biochem. et Biophys. Acta, 1097:107-143, Page, M.J. and Sydenham, M.A. 1991, Biotechnology Vol. 9:64-68). Cells grown in increasing concentrations of MTX develop resistance to the drug by overproducing the target enzyme, DHFR, as a result of amplification of the DHFR gene. If a second gene is linked to the DHFR gene it is usually co-amplified and over-expressed. It is state of the art to develop cell lines carrying more than 1,000 copies of the genes. Subsequently, when the methotrexate is withdrawn, cell lines contain the amplified gene integrated into the chromosome(s).

Plasmid pC1 contains for the expression of the gene of interest a strong promoter of the long terminal repeat (LTR) of the Rouse Sarcoma Virus (Cullen, *et al.*, Molecular and Cellular Biology, March 1985:438-4470) plus a fragment isolated from the enhancer of the immediate early gene of human cytomegalovirus (CMV) (Boshart *et al.*, *Cell 41*:521-530, 1985). Downstream of the promoter are the following single restriction enzyme cleavage sites that allow the integration of the genes: BamHI, Pvull, and Nrul. Behind these cloning sites the plasmid contains translational stop codons in all three reading frames followed by the 3′ intron and the polyadenylation site of the rat preproinsulin gene. Other high efficient promoters can also be used for the expression, e.g., the human β-actin promoter, the SV40 early or late promoters or the long terminal

repeats from other retroviruses, e.g., HIV and HTLVI. For the polyadenylation of the mRNA other signals, e.g., from the human growth hormone or globin genes can be used as well.

Stable cell lines carrying the gene of interest integrated into the chromosomes can also be selected upon co-transfection with a selectable marker such as gpt, G418 or hygromycin. It is advantageous to use more than one selectable marker in the beginning, e.g., G418 plus methotrexate.

The plasmid pC1 is digested with the restriction enzyme BamHI and then dephosphorylated using calf intestinal phosphates by procedures known in the art. The vector is then isolated from a 1% agarose gel.

The *M. jannaschii* protein-encoding sequence is is amplified using PCR oligonucleotide primers as described above. An efficient signal for initiation of translation in eukaryotic cells, as described by Kozak, M., J. Mol. Biol. 196:947-950 (1987) is appropriately located in the vector portion of the construct. The amplified fragments are isolated from a 1% agarose gel as described above and then digested with the endonucleases BamHI and Asp718 and then purified again on a 1% agarose gel.

The isolated fragment and the dephosphorylated vector are then ligated with T4 DNA ligase. *E. coli* HB101 cells are then transformed and bacteria identified that contained the plasmid pC1 inserted in the correct orientation using the restriction enzyme BamHI. The sequence of the inserted gene is confirmed by DNA sequencing.

Transfection of CHO-DHFR-cells

Chinese hamster ovary cells lacking an active DHFR enzyme are used for transfection. 5 μg of the expression plasmid C1 are cotransfected with 0.5 μg of the plasmid pSVneo using the lipofecting method (Felgner *et al., supra*). The plasmid pSV2-neo contains a dominant selectable marker, the gene neo from Tn5 encoding an enzyme that confers resistance to a group of antibiotics including G418. The cells are seeded in alpha minus MEM supplemented with 1 mg/ml

G418. After 2 days, the cells are trypsinized and seeded in hybridoma cloning plates (Greiner, Germany) and cultivated from 10-14 days. After this period, single clones are trypsinized and then seeded in 6-well petri dishes using different concentrations of methotrexate (25 nM, 50 nM, 100 nM, 200 nM, 400 nM). Clones growing at the highest concentrations of methotrexate are then transferred to new 6-well plates containing even higher concentrations of methotrexate (500 nM, 1 μM, 2 μM, 5 μM). The same procedure is repeated until clones grow at a concentration of 100 μM.

The expression of the desired gene product is analyzed by Western blot analysis and SDS-PAGE.

Example 7

Production of an Antibody to a Methanococcus jannaschii Protein

Substantially pure *M. jannaschii* protein or polypeptide is isolated from the transfected or transformed cells described above using an art-known method. The protein can also be chemically synthesized. Concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows:

Monoclonal Antibody Production by Hybridoma Fusion

Monoclonal antibody to epitopes of any of the peptides identified and isolated as described can be prepared from murine hybridomas according to the classical method of Kohler, G. and Milstein, C., *Nature 256*:495 (1975) or modifications of the methods thereof. Briefly, a mouse is repetitively inoculated with a few micrograms of the selected protein over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma

cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, E., *Meth. Enzymol. 70*:419 (1980), and modified methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. *et al.* Basic Methods in Molecular Biology Elsevier, New York. Section 21-2 (1989).

Polyclonal Antibody Production by Immunization

Polyclonal antiserum containing antibodies to heterogenous epitopes of a single protein can be prepared by immunizing suitable animals with the expressed protein described above, which can be unmodified or modified to enhance immunogenicity. Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. For example, small molecules tend to be less immunogenic than other molecules and may require the use of carriers and adjuvant. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. An effective immunization protocol for rabbits can be found in Vaitukaitis, J. *et al.*, *J. Clin. Endocrinol. Metab. 33*:988-991 (1971).

Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall (*See* Ouchterlony, O. *et al.*, Chap. 19 in: *Handbook of Experimental Immunology*, Wier, D., ed, Blackwell (1973)). Plateau

concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 $_{\mu}$M). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: *Manual of Clinical Immunology*, second edition, Rose and Friedman, (eds.), Amer. Soc. For Microbio., Washington, D.C. (1980).

Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample.

TTTTTATTTTATTATAAAATTCAAAAAATATCTTATCGTATTATAGAAAGATTTGTAATA
AACTCATTATAATAGTGAAATCTTACTTCGAAAATTTCTAACCTTGGCTGAACTTTGTTG
ATTAAGTTCAGGATAAACAAAAAATAAAAAGAACAATGATTTTAAACTCACTATCAGTGT
AGAGATTGGCATTAACTATTTATTTGTATTTATCTATCATACTGAGAGTTTTTTTATTTT
CTTTTATTGCTTTATTGATTTTTCTTTGAATGATTCTAGTACTATTTTCTCATAAGGAAA
AATGTTTGGTTTGTCATCTTAAATTTAAGTGATTTGATAAATTATAATTATCCCAACTTA
AACTGTAAATGAACTACAATATCCTTTCCTTTTTGTTTAAGTTCTATATCTTTTATTTTT
GAACAATTTCCACAGAATTCTTTTCTTAATATGTTTTTTATGTATCGGCATAAAGATTCT
TTGATTATTGCATCGTTTATATCATACCAAATTGCATAATTTTTGAGTTCGAATTCAAAA
TTTGGCTTTTTACTCTTCATTACCTCATATATTTCTTTAATAATATTGTCCCAGTTAGGT
TTAATATATTCTTCATTTAATAAGTTTTTATCGATATATTTTTCAATATTTTCTTTCTGT
GAGTCAAAACCATTTTTCTTTCGTATTATTTTATAGATTTTATACCCTCTTCCTAAAATC
CTATGGTTAGTGGATTGTATTTTTATCCATGTGTTAATTTCACGAATGGGTAGTCGTTGG
TCTATTACATAGATTTTTCCATTTATCTCAATAGCTGCAGCAACATGCATAGGATGAGTT
ACTAAATATAAGTTATAGTTGGGAAACAAATTCGAAAGTAAAGCAATAGTTAGTTTAGCA
TAATCTCTACATACTGCTTTTTTGTATTTTAGAATTTTGGACACTTTAATATCATAACAT
AAAGTATCATGCAGCATAGCGATTGCTGATACAAAATTTCCATTGCATTGAAACAAATAT
TTTACTATCAATGTAAGTATTAACACCCAAAGTCCTGAAATTATACTAATAATTATAAGA
TTTTCACTGAAAATATACAATGTGAAGATACTTACTAACAACAACATGCTAATGTTGTTT
AAGTATTGTGAATATTTTGATGGCAATGATATTAATACTAAGAGAGCCACTACTGCAGAT
ATTACCCATAGTATCAATACCATAATATCATTGATTATAATCTCAAAACCTATTATCAAT
AACAATACCATAAATAACAATACCACACCATATAACATAGCCGCAATAACATAATAAATT
AAAGAATCTGCCGCTCTTTCCATCCAATATCTAATATTAGTTTCTTGCCATTCCAAAATA
TTATTTAAAGTTTCAACAATTGAATTTTCCCATAACTGTTCAGACAGTTTTTTTATTTCG
TTACTATAAATTTCTTTTAGAGAAGGAATACTTAAAAAGTGTGACAACTT

While the present invention has been described in some detail for purposes of clarity and understanding, one skilled in the art will appreciate that various changes in form and detail can be made without departing from the true scope of the invention.

All patents, patent applications and publications recited herein are hereby incorporated by reference.

What Is Claimed Is:

1. An isolated nucleic acid molecule comprising a polynucleotide having a nucleotide sequence at least 95% identical to a sequence selected from the group consisting of:

(a) a nucleotide sequence of an open reading frame depicted in Table 2(a) or 3;

(b) a nucleotide sequence of an open reading frame depicted in Table 2(a) or 3, but minus the codon for the N-terminal methionine residue, if present; and

(c) a nucleotide sequence complementary to any of the nucleotide sequences in (a) or (b).

2. An isolated nucleic acid molecule comprising a polynucleotide having a nucleotide sequence 100% identical to a sequence in (a), (b) or (c) of claim 1.

3. An isolated nucleic acid molecule comprising a polynucleotide that hybridizes under stringent hybridization conditions to the nucleic acid molecule of claim 2.

4. An isolated nucleic acid molecule comprising a polynucleotide that encodes the amino acid sequence of an epitope-bearing portion of the *M. jannaschii* protein encoded by an open reading frame depicted in Table 2(a) or 3.

5. A method of making a recombinant vector comprising inserting the isolated nucleic acid molecule of claim 1 into a vector.

6. A recombinant vector produced by the method of claim 5.

7. A method of making a recombinant host cell comprising introducing the recombinant vector of claim 6 into a host cell.

8. A recombinant host cell produced by the method of claim 7.

9. A recombinant method for producing a *M. jannaschii* polypeptide, comprising culturing the recombinant host cell of claim 8 under conditions such that said polypeptide is expressed and recovering said polypeptide.

10. An isolated polypeptide having an amino acid sequence at least 95% identical to the amino acid sequence selected from the group consisting of:

(a) an amino acid sequence encoded by a *M. jannaschii* open reading frame depicted in Table 2(a) or 3; and

(b) an amino acid sequence encoded by a *M. jannaschii* open reading frame depicted in Table 2(a) or 3, but lacking the N-terminal methionine residue.

11. An isolated polypeptide, wherein said amino acid sequence is 100% identical to a sequence in (a) or (b) of claim 10.

12. An isolated antibody that binds specifically to the polypeptide of claim 11.

13. Computer readable medium having recorded thereon the nucleotide sequence depicted in SEQ ID NO:1, 2, or 3, or a nucleotide sequence at least 99.9% identical thereto.

14. Computer readable medium having recorded thereon the nucleotide sequence of at least one *M. jannaschii* open reading frame depicted in Table 2(a) or 3 or its complement.

15. The computer readable medium of claim 13, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

16. The computer readable medium of claim 14, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

17. A computer-based system for identifying fragments of the *M. jannaschii* genome that are homologous to target nucleotide sequences, comprising:

(a) a data storage means comprising the nucleotide sequence of SEQ ID NO:1, 2, or 3, or a nucleotide sequence at least 99.9% identical thereto;

(b) a search means for comparing a target sequence to said nucleotide sequence of said data storage means of step (a) to identify a homologous sequence, and

(c) a retrieval means for obtaining said homologous sequence of step (b).

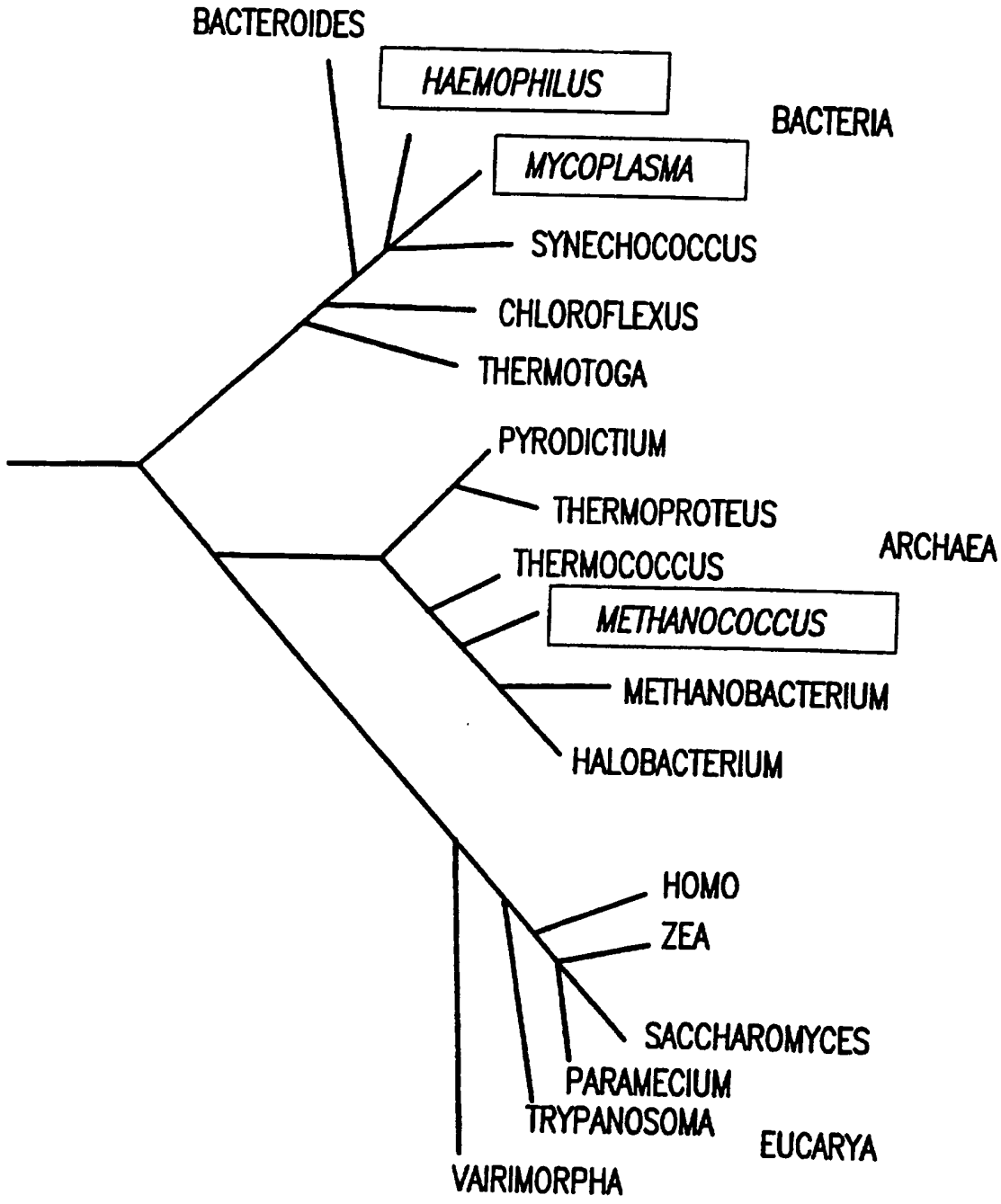

BACTEROIDES
HAEMOPHILUS
BACTERIA
MYCOPLASMA
SYNECHOCOCCUS
CHLOROFLEXUS
THERMOTOGA
PYRODICTIUM
THERMOPROTEUS
ARCHAEA
THERMOCOCCUS
METHANOCOCCUS
METHANOBACTERIUM
HALOBACTERIUM
HOMO
ZEA
SACCHAROMYCES
PARAMECIUM
TRYPANOSOMA
EUCARYA
VAIRIMORPHA

FIG.1

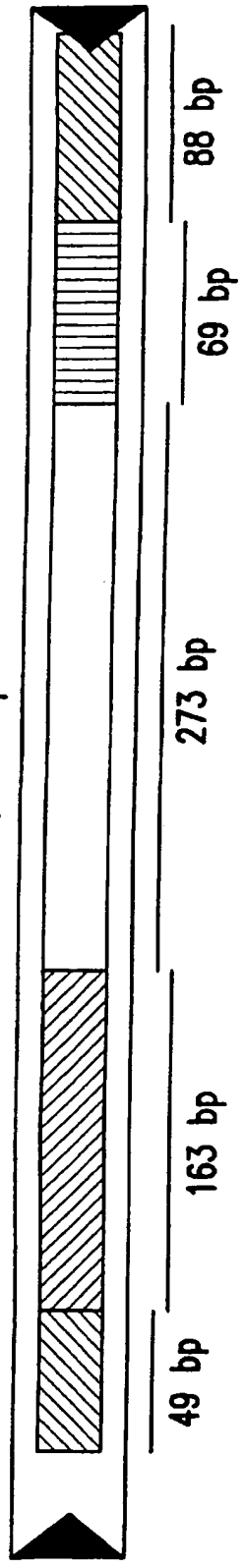


FIG.2A

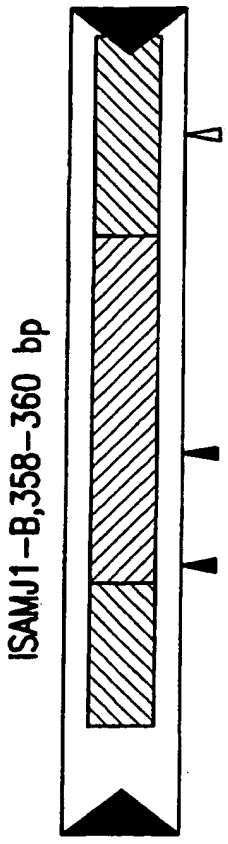


FIG.2B

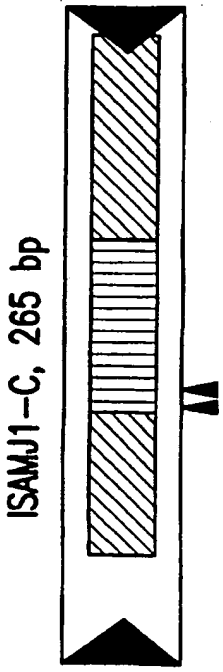


FIG.2C

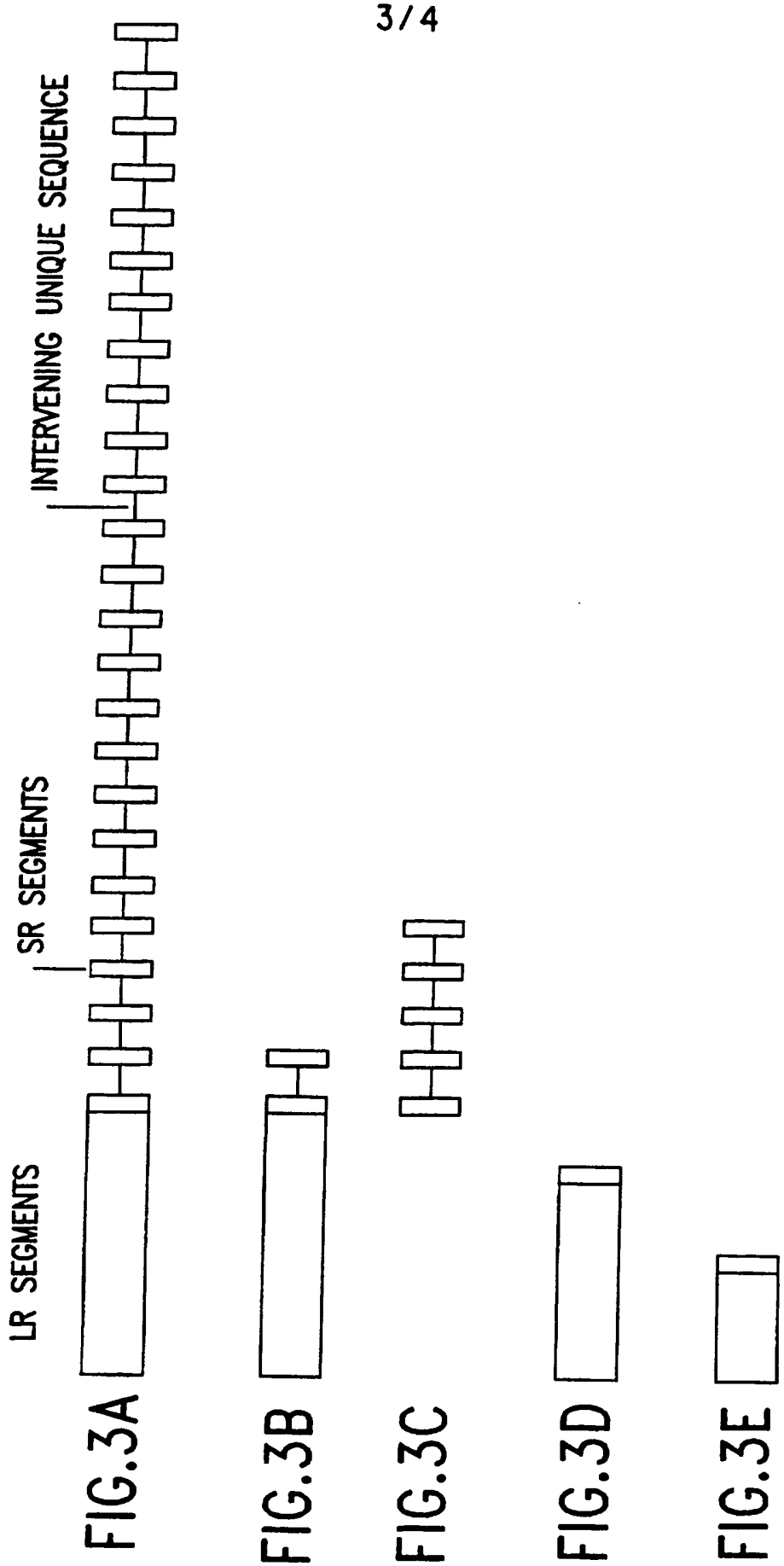


FIG.3A FIG.3B FIG.3C FIG.3D FIG.3E

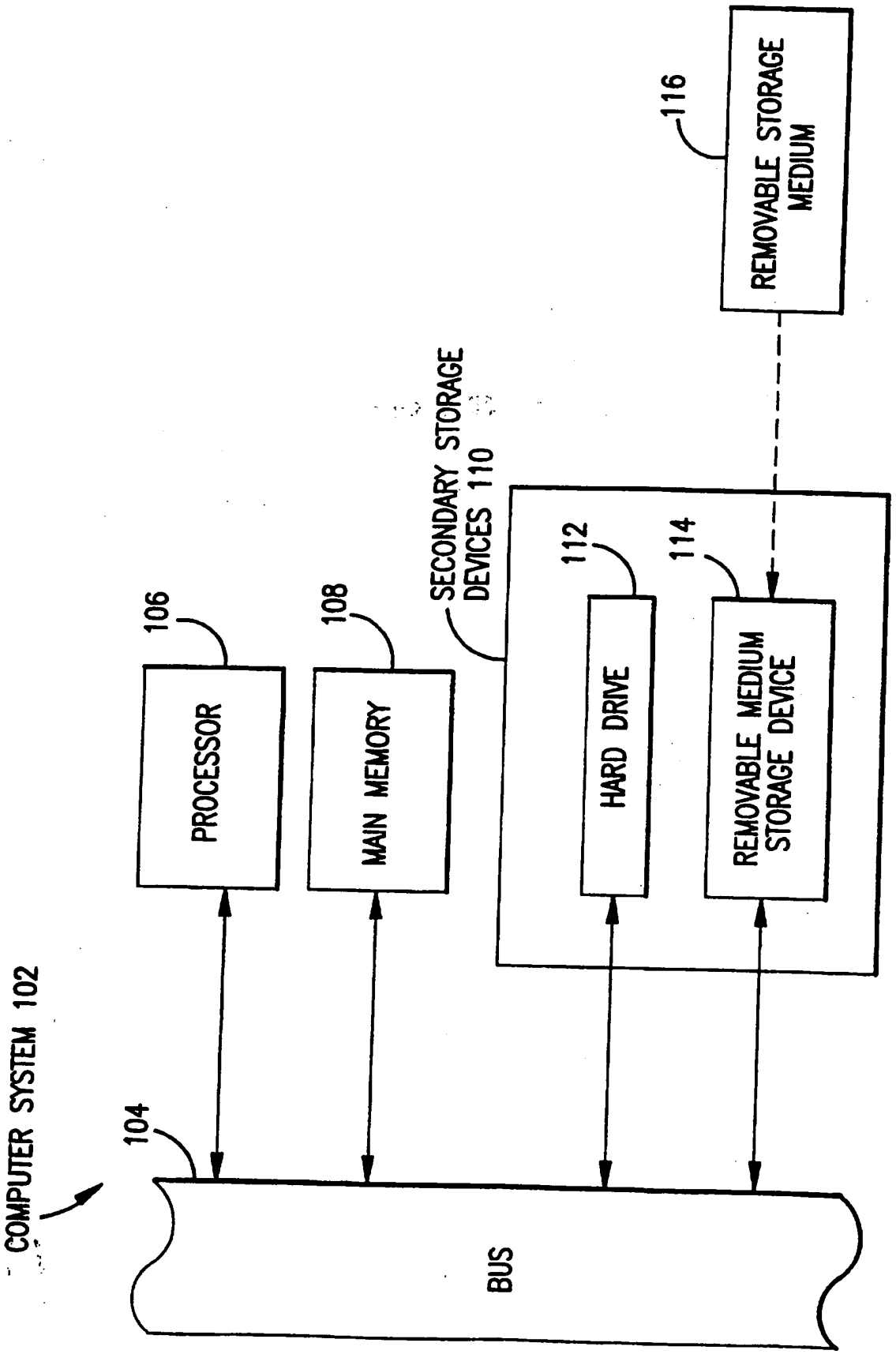

COMPUTER SYSTEM 102
104
BUS
PROCESSOR
106
MAIN MEMORY
108
SECONDARY STORAGE DEVICES 110
HARD DRIVE
112
REMOVABLE MEDIUM STORAGE DEVICE
114
REMOVABLE STORAGE MEDIUM
116

FIG.4